Global Coastal Groundwater Dynamics and Predictive Modeling Based on Harmonized In Situ Groundwater Level Observations

Dissertation

zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und
Naturwissenschaften Fachbereich Erdsystemwissenschaften
der Universität Hamburg

vorgelegt von
Annika Nolte

Hamburg
September 08, 2026

Fachbereich Erdsystemwissenschaften

Datum der Disputation:	28.05.2026
Gutachter der Dissertation:	Prof. Dr. Jens Hartmann Dr. Laurens Bouwer
Zusammensetzung der Prüfungskommission:	Prof. Dr. Jens Hartmann Dr. Laurens Bouwer Prof. Dr. Johanna Baehr Prof. Dr. Jürgen Böhner Dr. Stefan Hagemann
Vorsitzender des Fach-Promotionsausschusses Erdsystemwissenschaften:	Prof. Dr. Dirk Notz
Dekan der Fakultät MIN:	Prof. Dr. Ing. Norbert Ritter

Abstract

Coastal aquifers are essential for water security and ecosystems. Yet, at continental to global scales, groundwater availability and vulnerability remain difficult to assess consistently. The evidence base is dominated by local case studies that are only partly comparable and by spatially continuous, top-down estimates whose results depend strongly on model assumptions. At the same time, in situ observations from groundwater monitoring are fragmented, unevenly distributed, and often not standardized. This thesis addresses this gap through a bottom-up approach that harmonizes, compiles, and integrates groundwater level (GWL) data from many monitoring locations worldwide, analyzes these data at large scale, and uses them for modeling groundwater dynamics. A key output is the Coastal GroundWater Level (CGWL) dataset, comprising around half a million coastal monitoring locations. The overarching aim of this thesis is to advance understanding of coastal groundwater dynamics and to improve their prediction at global scale. To this end, it uses the CGWL dataset to deliver three complementary, observation-based contributions.

First, this thesis derives head- and gradient-based diagnostics to characterize land–sea exchange conditions and their decadal evolution as indicators of seawater intrusion (SWI) susceptibility. SWI occurs when seawater enters coastal groundwater, increases salinity, and threatens freshwater availability. Across monitored coastlines, roughly one-third of coastal areas have gradients indicative of reduced freshwater discharge potential and higher SWI susceptibility, while 21% show significant GWL changes of at least ± 0.1 m yr^{-1} between 1990 and 2024. Close to the shoreline, head changes tend to occur where SWI susceptibility is already highest, with potential to shift land–sea gradients over time, including within a decade. In addition, changes are more pronounced for deeper groundwater and in arid regions, and they are not confined to urban centers. The results provide globally comparable evidence to support risk screening and monitoring prioritization and to benchmark and inform large-scale models.

Second, four recurring archetypal dynamic regimes in thousands of GWL hydrographs from coastal regions of different continents and climate zones are identified. The derived typology serves as a scalable reference for global comparison and enables an assessment of hydrogeological similarity across spatial scales, from local variability within regions to broad global patterns. Explainable machine learning (XAI) indicates a first-order partitioning of dynamic regimes by groundwater depth and topography, with further differentiation by climate and soil/lithology. The analysis of the extent to which regimes can be explained from globally available site descriptors shows that current proxy datasets

represent key controls only partly, which sets realistic expectations for transferring or regionalizing observed dynamics based on these datasets in large-scale modeling approaches.

Third, the feasibility and current limits of scalable deep learning for weekly-ahead GWL prediction is demonstrated by benchmarking multi-site models across 1,800 wells from nine regions worldwide under realistic monitoring constraints and missing or unreliable site descriptors. It is shown that high predictive skill can be achieved in data-rich monitoring networks using only dynamic climatic inputs and site identifiers, and that pooled multi-region training can serve as a scalable default. However, predictive skill is strongly constrained by measurement frequency and measurement regularity, and adding data from data-rich regions does not improve prediction in data-scarce regions, revealing current limits to generalization. Learned internal site representations provide a compact, scalable alternative to external site descriptors and enable post hoc analyses of similarity between wells and model behavior, but they require independent context for physical interpretation and validation.

Overall, this thesis converts fragmented coastal monitoring into globally comparable, in situ evidence that enables large-scale assessment of coastal groundwater systems, particularly their state, dynamics, and predictability. While the global-scale perspective precludes unambiguous attribution of the drivers of observed quantitative changes in groundwater and reflects the monitoring footprint and data and metadata limitations, the thesis clarifies what can – and cannot yet – be summarized and generalized about coastal groundwater at the global scale. In doing so, it underscores the value of globally comparable, yet locally faithful evidence from groundwater monitoring as a basis for observation-based diagnostics, model evaluation and development, and targeted improvements of the monitoring needed to support them.

Zusammenfassung

Küstennahe Grundwasserleiter sind weltweit essenziell für eine sichere Wasserversorgung und Ökosysteme. Gleichzeitig lassen sich die Verfügbarkeit und Vulnerabilität von Grundwasser auf kontinentalen bis globalen Skalen bislang nur schwer konsistent beurteilen. Die Evidenzbasis wird einerseits von lokalen Fallstudien dominiert, die nur bedingt vergleichbar sind, und andererseits von räumlich flächendeckenden Top-down-Schätzungen, deren Aussagen stark von Modellannahmen abhängen. Zugleich sind In situ Beobachtungen aus dem Grundwassermonitoring fragmentiert, räumlich ungleich verteilt und häufig nicht standardisiert. Diese Dissertation adressiert diese Lücke mit einem Bottom-up-Ansatz, bei dem Grundwasserstandsdaten aus vielen Messstellen weltweit harmonisiert, zusammengeführt, analysiert und zur Modellierung der Grundwasserdynamik genutzt werden. Ein zentrales Ergebnis ist der Coastal GroundWater Level (CGWL)-Datensatz, der rund eine halbe Million küstennaher Messstellen umfasst. Das übergeordnete Ziel dieser Dissertation ist es, zu einem besseren Verständnis der Dynamik und Entwicklung küstennaher Grundwassersysteme beizutragen und deren Vorhersage auf globaler Skala zu verbessern. Dazu liefert diese Dissertation auf Basis des CGWL-Datensatzes drei komplementäre, messdatenbasierte Beiträge.

Erstens werden Kennwerte aus hydraulischen Höhen (Grundwasserstand bzw. piezometrische Höhe) und Land–Meer-hydraulischen Gradienten abgeleitet, um die Austauschverhältnisse an der Land–Meer-Grenze sowie deren dekadische Entwicklung als Indikatoren für eine erhöhte Suszeptibilität gegenüber Salzwasserintrusion (SWI) zu beschreiben. Bei SWI dringt Meerwasser in küstennahes Grundwasser ein, erhöht dort die Salinität und gefährdet die Verfügbarkeit der Süßwasserressource. Entlang der durch Messdaten abgedeckten Küstenabschnitte weisen etwa ein Drittel Gradienten auf, die auf ein vermindertes Potenzial für Grundwasserabfluss zum Meer und eine erhöhte SWI-Suszeptibilität hindeuten. Zugleich zeigen 21% signifikante Änderungen der Grundwasserstände von mindestens $\pm 0{,}1$ m a^{-1} im Zeitraum 1990–2024. Küstennahe Änderungen treten dabei häufig dort auf, wo die SWI-Suszeptibilität bereits am höchsten ist, und können Land–Meer-Gradienten über die Zeit verschieben, auch innerhalb eines Jahrzehnts. Zudem sind Änderungen bei tieferen Grundwasserständen und in ariden Regionen ausgeprägter und nicht auf urbane Zentren beschränkt. Die Ergebnisse liefern global vergleichbare Evidenz für Risikoeinordnung und Priorisierung des Monitorings sowie beobachtungsbasierte Referenzen zur Modellüberprüfung und -weiterentwicklung großskaliger Modelle.

Zweitens identifiziert die Dissertation vier wiederkehrende archetypische Dynamikregime

in Tausenden Grundwasserstandsganglinien aus Küstenregionen unterschiedlicher Kontinente und Klimazonen. Die daraus abgeleitete Typologie dient als skalierbare Referenz für globale Vergleiche und ermöglicht eine Bewertung hydrogeologischer Ähnlichkeit über verschiedene räumliche Maßstäbe hinweg, von lokaler Variabilität innerhalb von Regionen bis zu großskaligen globalen Mustern. Erklärbares Machine Learning (XAI) zeigt eine erste, grobe Trennung der Regime entlang von Grundwassertiefe und Topographie, mit weiterer Differenzierung durch Klima sowie Boden/Lithologie. Die Analyse der Erklärbarkeit der Regime anhand global verfügbarer Standortmerkmale zeigt, dass derzeitige Proxy-Datensätze zentrale Kontrollgrößen nur teilweise abbilden, was realistische Erwartungen an die Übertragbarkeit und Regionalisierung beobachteter Dynamiken auf Basis dieser Datensätze in großskaligen Modellen setzt.

Drittens wird die Machbarkeit und aktuelle Grenzen skalierbarer Deep-Learning-Modellierung für wöchentliche Grundwasserstandsvorhersagen gezeigt, indem messstellenübergreifend trainierte Modelle an 1.800 Messstellen aus neun Regionen weltweit unter realistischen Monitoringbedingungen sowie bei fehlenden oder unzuverlässigen Standortmerkmalen systematisch bewertet werden. In datenreichen Messnetzen lässt sich bereits mit dynamischen klimatischen Eingangsgrößen und Standortkennungen eine hohe Vorhersagegüte erzielen, und ein gemeinsam über mehrere Regionen trainiertes Modell kann dort als skalierbarer Standard dienen. Die Vorhersagegüte wird jedoch stark durch Messfrequenz und Messregelmäßigkeit begrenzt, und zusätzliche Trainingsdaten aus datenreichen Regionen verbessern Vorhersagen in datensarmen Regionen nicht, was Grenzen der Generalisierbarkeit sichtbar macht. Vom Modell gelernte interne Standortprofile bieten eine kompakte, skalierbare Alternative zu externen Standortmerkmalen und ermöglichen nachgelagerte Analysen von Ähnlichkeit zwischen Messstellen und Modellverhalten, erfordern für physikalische Interpretation und Validierung jedoch unabhängigen Kontext.

Insgesamt überführt diese Dissertation fragmentiertes küstennahes Monitoring in global vergleichbare, beobachtungsbasierte Evidenz, die eine großskalige Beurteilung küstennaher Grundwassersysteme ermöglicht, insbesondere ihres Zustands, ihrer Dynamik und ihrer Vorhersagbarkeit. Obwohl der globale Maßstab eine eindeutige Zuordnung der Treiber mengenmäßiger Veränderungen und Dynamiken im Grundwasser ausschließt und die Ergebnisse die räumliche Abdeckung des Monitorings sowie Einschränkungen in Daten und Metadaten widerspiegeln, zeigt die Dissertation, was sich über küstennahes Grundwasser auf globaler Skala zusammenfassen und verallgemeinern lässt und was derzeit noch nicht. Damit unterstreicht sie den Wert global vergleichbarer, zugleich lokal verankerter Evidenz aus dem Grundwassermonitoring als Grundlage für messdatenbasierte Diagnostiken,

Modellbewertung und Weiterentwicklung sowie für eine gezielte Verbesserung des dafür notwendigen Monitorings.

Acknowledgement

Looking back, this PhD has been a rewarding journey, even though it was challenging at times. I was fortunate to be supported by many people – both in my research and on a personal level.

First, I want to thank my supervisors, *Steffen Bender* and *Jens Hartmann*. Thank you for making this project possible for me and for your scientific guidance and steady support. In the final stage of this project, I am also grateful to *Laurens Bouwer* for his co-supervision and valuable advice over the last few months.

Thank you, *Robert Reinecke*, *Benedikt Heudorfer*, *Ezra Haaf*, and *Nils Moosdorf*, for your excellent collaboration on individual projects — not only as co-authors, but also for your empathy and genuine interest in my work.

I also want to thank my project colleague *Stefan Baltruschat* for insightful and enjoyable discussions on database development and statistics, and for his collegial support. Thank you, *Christina Pop*, for the time we spent together during breaks at the Chilehaus, which was always a welcome opportunity for exchange and a refreshing pause for me. Thanks also to all the kind and supportive people I had the pleasure to meet and exchange with at GERICS and in my former working group at UHH, particularly *Thorben Amann*, for his technical support. I would also like to thank my new colleagues at the Data Science Center for the great working atmosphere and personal support.

Many thanks also go to my friends and current and former flatmates, especially *Lina*, *Sabine*, *Jasmin*, *Johanna*, *Gregor*, and *Nino*, as well as to all the wonderful people who have stayed close over the years since my student days in Braunschweig, and to friends in Süstedt, Hamburg, and Bremen. Thank you for your encouragement and for keeping me grounded.

Finally, I am deeply grateful to my parents, *Vera* and *Heiner*, and my sister, *Maybrit*. Thank you for supporting me in all my endeavors and for helping me build confidence and resilience that also carried me through this PhD. And thank you, *Jannik*, for your endless patience, your belief in me, and for always making sure I kept my balance throughout these years.

Table of contents

Abstract v

Zusammenfassung vii

Acknowledgement x

Abbreviations and acronyms xvi

List of figures xviii

List of tables xxi

List of publications xxii

1 Introduction 1
- 1.1 Background and problem framing . . . 1
- 1.2 Large-scale groundwater assessments . . . 4
  - 1.2.1 Observational basis . . . 4
  - 1.2.2 Modeling approaches . . . 6
  - 1.2.3 Knowledge trade-off and strategies . . . 8
- 1.3 Research gaps and directions . . . 9
- 1.4 Thesis outline and research questions . . . 11

2 The Coastal GroundWater Level (CGWL) dataset 13

3 Coastal groundwater level trends reveal global susceptibility to seawater intrusion 18
- 3.1 Abstract . . . 18
- 3.2 Introduction . . . 18
- 3.3 Susceptible hydroclimatic conditions . . . 20
- 3.4 Prevalence of trends and sensitivity to record length . . . 22
- 3.5 Trends in rural and urban coastal regions . . . 24
- 3.6 Identifying hotspots of seawater intrusion susceptibility . . . 25
- 3.7 Discussion . . . 26
- 3.8 Conclusions . . . 29
- 3.9 Methods . . . 30
  - 3.9.1 Groundwater hydraulic gradients . . . 30
  - 3.9.2 Trend analysis . . . 30

3.9.3 Hydroclimatic classification and hotspot extrapolation 31
3.9.4 Preparing secondary datasets for contextual analyses 32
3.9.5 Limitations 32

4 Disentangling coastal groundwater level dynamics in a global dataset 35
4.1 Abstract 35
4.2 Introduction 35
4.3 Data and methods 38
4.3.1 GWL time series and preprocessing 38
4.3.2 Indices 40
4.3.3 Cluster analysis 40
4.3.4 Attribute data 41
4.3.5 Random forest model 45
4.3.6 SHAP analysis 46
4.4 Results 46
4.4.1 Identification of groundwater level dynamic patterns 46
4.4.2 Classification of groundwater level dynamic patterns 49
4.4.3 Model explanations from SHAP values 50
4.5 Discussion 54
4.5.1 Hydrogeological similarity and scaling effect 54
4.5.2 Cause-and-effect relationships 57
4.5.3 Case study 59
4.6 Conclusions 62

5 Multi-site deep learning for groundwater level prediction across global datasets: toward scalable applications under data scarcity 64
5.1 Abstract 64
5.2 Introduction 64
5.3 Study area and data 68
5.4 Methods 69
5.4.1 Scope and experimental design 69
5.4.2 Standard LSTM 71
5.4.3 LSTM with embeddings 74
5.4.4 Exploring learned representations through embeddings 74
5.5 Results and discussion 76
5.5.1 Model performance across regions 76
5.5.2 Superior performance of entity-aware models 78
5.5.3 Global models match regional performance 79

5.5.4 Data quantity and quality limitations 81
5.5.5 Insights into spatial representation through embeddings 84
5.6 Conclusions 88

6 Synthesis **90**
6.1 Land–sea exchange and susceptibility to seawater intrusion 90
6.2 Groundwater dynamic regimes and explanations 92
6.3 Predictive modeling of groundwater levels at the global scale 93
6.4 Conclusions and outlook 94

Bibliography **97**

A The Coastal GroundWater Level (CGWL) dataset **119**
A.1 Data acquisition and preprocessing 119
A.2 Limitations in differentiating aquifers at large scales 121
A.3 Data aggregation 123
A.4 CGWL dataset structure 124
A.5 Spatiotemporal coverage 130
A.6 Source-specific details on data acquisition and preprocessing 133

B Coastal groundwater level trends reveal global susceptibility to seawater intrusion **137**
B.1 Time series selection and filtering criteria for trend analysis 137
B.2 Spatial availability of trend results 138
B.3 Sensitivity of trend estimates to record length 138
B.4 Sensitivity of results to spatial aggregation 141
B.5 Supporting trend results 144

C Disentangling coastal groundwater level dynamics in a global dataset **147**
C.1 Time series outlier detection with DBSCAN 147
C.2 Indices 148
C.3 Clustering 151
C.4 Random forest classification 156
C.5 SHAP 157

D Multi-site deep learning for groundwater level prediction across global datasets: toward scalable applications under data scarcity **164**
D.1 Descriptive statistics of groundwater and climate data 165
D.2 Model hyperparameters 167

D.3 Model evaluation 167

Use of Electronic Tools **i**

Eidesstattliche Erklärung | Declaration on Oath **ii**

Abbreviations and acronyms

AI	aridity index
C1–C4	cluster 1–4 (classification labels used in Chapter 3)
CAMELS	Catchment Attributes and Meteorology for Large-sample Studies
CDF	cumulative distribution function
CGWL	Coastal GroundWater Level (dataset – Chapter 2)
CL1–CL4	cluster 1–4 (classification labels used in Chapter 4)
CNN	convolutional neural network
CoastalDEM	CoastalDEM digital elevation model
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DEM	digital elevation model
DL	deep learning
ECMWF	European Centre for Medium-Range Weather Forecasts
ENSO	El Niño–Southern Oscillation
ERA5	fifth generation ECMWF atmospheric reanalysis of the global climate
FEFLOW	Finite Element subsurface FLOW system
GGMN	Global Groundwater Monitoring Network
GLHYMPS	Global HYdrogeology MaPS
GRACE	Gravity Recovery and Climate Experiment
GWL	groundwater level
H3	H3 hexagonal spatial indexing system
HYBAS	HydroBASINS identifier
IGRAC	International Groundwater Resources Assessment Centre
IPCC	Intergovernmental Panel on Climate Change
KGE	Kling–Gupta efficiency
LSTM	long short-term memory
MAE	mean absolute error
ML	machine learning
MODFLOW	Modular Finite-Difference groundwater flow model
MSE	mean squared error
NMI	normalized mutual information

NSE	Nash–Sutcliffe efficiency
OBS	observations
PC	principal component
PCA	principal component analysis
RF	random forest
RMSE	root mean square error
RQ	research question
SDGs	Sustainable Development Goals
SGD	submarine groundwater discharge
SGR	submarine groundwater recharge
SHAP	SHapley Additive exPlanations
SIM	simulation
SPI	standardized precipitation index
SPEI	standardized precipitation evapotranspiration index
SRTM	Shuttle Radar Topography Mission
SWI	seawater intrusion
t-SNE	t-distributed Stochastic Neighbor Embedding
TWS	terrestrial water storage
WTD	water table depth

List of Figures

Figure 1	A perceptual model of drivers and limiters of coastal groundwater dynamics	3
Figure 2	Spatial coverage of the CGWL dataset.	15
Figure 3	Comparison of water table depths in the CGWL dataset with previous global estimates	16
Figure 4	Global patterns of hydroclimatic conditions indicating susceptibility of coastal groundwater systems to seawater intrusion.	22
Figure 5	Spatial patterns of recent groundwater level trends across urban and rural global coastal zones, grouped by IPCC regions.	23
Figure 6	Share of wells with upward and downward trends by distance from the coast, stratified by seawater intrusion susceptibility.	25
Figure 7	Spatial patterns of seawater intrusion susceptibility hotspots and mapped projected recharge change and groundwater withdrawals.	28
Figure 8	Assignment of groundwater dynamic regimes (clusters) to well locations at the global scale and regionally.	47
Figure 9	Typical annual groundwater hydrographs for clusters across coastal regions.	48
Figure 10	Example groundwater time series illustrating components of groundwater level dynamics in Europe.	52
Figure 11	Performance of the random forest classification: evaluation metrics and normalized confusion matrix.	53
Figure 12	Groundwater level dynamic patterns in the German federal state Schleswig-Holstein.	60
Figure 13	Focus regions and well locations in regional and global GWL modeling.	70
Figure 14	LSTM model architecture in three variants.	72
Figure 15	Performance of three model variants across nine focus regions, showing well-level test-score distributions and run-to-run variability.	77
Figure 16	Comparison of global and regional model performance.	80
Figure 17	Example groundwater level time series in the test period of each region showing observed and simulated values.	82
Figure 18	Geographic structure in the learned embedding space.	85
Figure 19	Groundwater dynamics structure in the learned embedding space.	86

Figure 20 Model performance differences for inlier and outlier wells in the embedding space. 87

Figure A1 Comparison of well land-surface elevations from groundwater datasets and digital elevation models. 121
Figure A2 Evaluation of a global near-surface proxy for aquifer state using the CGWL dataset 122
Figure A3 Within-unit spread in median groundwater depth across grids and basins, and how many wells contribute to each unit. . . . 124
Figure A4 Structure of the CGWL dataset. 126
Figure A5 Spatial distribution of grid-level observations in the CGWL dataset. 131
Figure A6 Temporal distribution of groundwater level observations in the CGWL dataset. 131
Figure A7 Temporal statistics for groundwater level time series in the CGWL dataset at three temporal aggregation levels. 132
Figure B1 Temporal distribution of available years in groundwater level time series used for trend analysis 137
Figure B2 Data availability for 9- and 19-year groundwater level trends across IPCC coastal regions. 138
Figure B3 Consistency of 9- and 19-year groundwater level trend slopes across grid cells and IPCC regions. 140
Figure B4 Sensitivity of key diagnostics for seawater intrusion susceptibility to the choice of spatial aggregation unit. 142
Figure C1 Examples of groundwater level time series outlier detection with DBSCAN and illustration of its limitations. 147
Figure C2 Comparison of groundwater hydrograph index value distributions from this study with those previously published. 149
Figure C3 Correlation matrix of the index values calculated in this study. 150
Figure C4 Principal component analysis with the index dataset. 151
Figure C5 Clustering evaluation metrics for model selection and assessment of cluster separation quality. 151
Figure C6 Cluster assignment of groundwater level time series by different algorithms and their comparative distribution. 152
Figure C7 Within-cluster variability of all indices used for clustering. . . 153

Figure C8 Within-cluster variability of all indices used for clustering and index values of four example wells from the case study. 154
Figure C9 Cluster assignments of wells in the Normandy, Paris Basin. . 155
Figure C10 Random forest hyperparameter tuning within a 5-fold cross-validation framework using the training data. 156
Figure C11 Absolute Spearman correlation of SHAP values for aggregated features. 157
Figure C12 Overall SHAP feature importance stacked for individual clusters for aggregated features. 158
Figure C13 Distribution of SHAP feature importance and feature effect for the 20 highest-ranked features of each cluster. 159
Figure C14 Explanations for the correct prediction of the CL1 exemplary well site from the case study. 160
Figure C15 Explanations for the prediction of the CL3 exemplary well site from the case study that is confused with CL2. 161
Figure C16 Explanations for the correct prediction of the CL3 exemplary well site from the case study. 162
Figure C17 Explanations for the prediction of the CL3 exemplary well site from the case study that is confused with CL4. 163
Figure D1 Boxplots showing groundwater level time series data quantity metrics for the studied regions. 165
Figure D2 Boxplots showing mean weekly precipitation and surface temperature values for the well locations in the studied regions. . 166
Figure D3 Maps of well-level model performance. 169
Figure D4 Learning curves of training and validation loss for all regions and model variants. 171

List of Tables

Table 1 List of sources from which groundwater level time series were obtained. 14
Table 2 List of attributes used to relate groundwater level dynamics to controls and processes. 43
Table 3 Qualitative description of key features driving cluster differentiation. 51
Table 4 Selection criteria and median groundwater levels for time series used in predictive modeling across regions. 68
Table 5 Summary statistics of model performance across regions and model variants. 76

Table A1 Metadata for the groundwater level time series. 125
Table A2 Metadata for the groundwater level time series statistics. . . . 126
Table A3 Metadata for the well locations. 127
Table A4 Metadata for the references. 128
Table A5 Metadata for the grids. 128
Table A6 Metadata for the basins. 128
Table A7 Metadata for the groundwater depth in grids. 129
Table A8 Metadata for the groundwater depth in basins. 129
Table A9 Metadata for the groundwater elevation in grids. 129
Table A10 Metadata for the groundwater elevation in basins. 130
Table A11 Overview of source-specific data acquisition and preprocessing in compiling the CGWL dataset 133
Table B1 IPCC-region-specific trend estimates combining 9- and 19-year windows. 144
Table B2 IPCC-region-specific 9-year trend estimates. 145
Table B3 IPCC-region-specific 19-year trend estimates. 146
Table C1 List of groundwater level dynamic typology and indices used in this study. 148
Table D1 Additional metrics for model evaluation across regions and model variants. 168

List of publications

This cumulative dissertation comprises three research articles that have been published in peer-reviewed journals or are currently under review. Together, these articles constitute the core scientific contributions of the dissertation. In addition, a data publication provides the global dataset used in the studies. The data publication and the three articles are presented in Chapters 2–5 of this thesis. They are listed below, together with additional related work to which the author contributed. For each article, the author's individual contributions are summarized using the CRediT (Contributor Roles Taxonomy) framework.

Data publication: **Chapter 2**

Nolte, **A.** (2026). Coastal GroundWater Level (CGWL) Dataset: [data set]. 10.5281/zenodo.15008065

Research articles:

Chapter 3

Nolte, **A.**, Bender, S., Hartmann, J., Baltruschat, S., Moosdorf, N., and Reinecke, R. (2026). "Coastal groundwater-level trends reveal global susceptibility to seawater intrusion". In: *Nature Water* 4, 673–682. 10.1038/s44221-026-00619-8

Author's contribution (AN): AN led the conceptualization and formal analysis, with contributions from R. Reinecke (RR). AN developed the methodology, supported by RR, N. Moosdorf, and S. Baltruschat. AN handled data curation, investigation, validation, and visualization. AN drafted the original manuscript with contributions from RR and revised it based on feedback from all co-authors.

Chapter 4

Nolte, **A.**, Haaf, E., Heudorfer, B., Bender, S., and Hartmann, J. (2024). "Disentangling coastal groundwater level dynamics in a global dataset". In: *Hydrology and Earth System Sciences* 28.5, 1215–1249. 10.5194/hess-28-1215-2024

Author's contribution (AN): AN developed the conceptualization and methodology, with contributions from E. Haaf and B. Heudorfer. AN handled data curation, software development, formal analysis, investigation, validation, and visualization. AN drafted the original manuscript and revised it based on feedback from all co-authors.

Chapter 5

Nolte, **A.**, Heudorfer, B., Bender, S., and Hartmann, J. (2025). "Multi-site deep learning for groundwater level prediction across global datasets: toward scalable applications under data scarcity". In: *Journal of Hydroinformatics* 27.10, 1632–1651. 10.2166/hydro.2025.095

Author's contribution (AN): AN developed the conceptualization, with contributions from B. Heudorfer (BH). AN and BH developed the methodology and software. AN handled data curation, formal analysis, investigation, validation, and visualization. AN drafted the original manuscript and revised it based on feedback from all co-authors.

Additional related work:

Panthi, J., Pradhanang, S. M., **Nolte**, **A.**, and Boving, T. B. (2022). "Saltwater intrusion into coastal aquifers in the contiguous United States - A systematic review of investigation approaches and monitoring networks". In: *Science of the Total Environment* 836, 155641. 10.1016/j.scitotenv.2022.155641

Author's contribution (AN): AN contributed to the research conceptualization led by J. Panthi (JP). AN and JP carried out the literature review, data collection, and formal analysis. AN contributed to manuscript writing and review.

Nolte, **A.**, Eley, M., Schöniger, M., Gwapedza, D., Tanner, J., Mantel, S. K., and Scheihing, K. (2021). "Hydrological modelling for assessing spatio–temporal groundwater recharge variations in the water–stressed Amathole Water Supply System, Eastern Cape, South Africa". In: *Hydrological Processes* 35.6, e14264. 10.1002/hyp.14264

Author's contribution (AN): AN conceptualized the study and developed the methodology together with M. Eley (ME) and M. Schöniger (MS). AN handled the data curation, formal analysis, investigation, validation, and visualization. AN drafted the original manuscript with contributions from K. Scheihing, ME, and MS, and revised it based on feedback from all co-authors.

1 Introduction

1.1 Background and problem framing

Freshwater is the source of all life, and groundwater is the Earth's largest resource of liquid freshwater (Aeschbach-Hertig and Gleeson 2012). Globally, groundwater supplies half of domestic water demand, sustains 38% of irrigated land, and various industrial processes rely on groundwater resources (United Nations 2022). As part of the hydrologic cycle, groundwater sustains streams, lakes, wetlands, and terrestrial and aquatic ecosystems (Alley et al. 1999). Therefore, groundwater is a common "pool resource" (Ostrom 1992) at the center of competition for water between different human and environmental needs. These tensions are especially pronounced in coastal regions, which are currently home to 2.86 billion people (Cosby et al. 2024) and are expected to face large population growth (Neumann et al. 2015). Along the land–sea interface, aquifers sustain valuable ecosystems and their services (Dyring et al. 2022; Luijendijk et al. 2020). Accordingly, ensuring the long-term sustainability of coastal groundwater resources is essential for coastal water security and for maintaining key environmental processes along coasts, with direct relevance to progress on multiple Sustainable Development Goals (SDGs) (Guppy et al. 2018; United Nations 2022).

Coastal groundwater resources are inherently dynamic. Groundwater systems comprise subsurface water stored and moving within a geologic framework and bounded by terrestrial (e.g., recharge, rivers and lakes, topography) and marine boundary conditions (e.g., sea level, tides, waves, storms) (Sophocleous 2002; Alley et al. 2002; Taniguchi et al. 2002). Aquifer properties such as permeability, thickness, heterogeneity, and the degree of confinement govern transmissivity, whether flow occurs mainly through the porous matrix or fractures, and how strongly the system connects to rivers, lakes, wetlands, and the coastal ocean (Bear 1979; Sophocleous 2002). In doing so, they act as key limiters that constrain the magnitude and rate of hydraulic head (i.e., the groundwater water level or pressure level observed in wells) responses to external forcing. Groundwater flow is driven by hydraulic gradients (Darcy 1856), and in coastal settings is further shaped by the land-sea boundary and the density contrast between fresh and saline water. These factors regulate freshwater discharge to the coastal ocean, seawater recirculation through the beach and nearshore aquifer, and the extent and position of the freshwater–saltwater transition zone (Taniguchi et al. 2002; Michael et al. 2005; Werner et al. 2013). Groundwater discharge conveys groundwater to surface waters, the coastal ocean, or wells, while groundwater recharge supplies inflow primarily via precipitation

infiltration (diffuse recharge) and via focused exchange with surface waters (localized recharge) (Alley et al. 2002; Vries and Simmers 2002).

Human activities and climate change increasingly reshape coastal groundwater regimes by altering the pathways and rates of recharge, discharge, and storage, and by shifting terrestrial and marine boundary conditions along the land–sea interface (Figure 1). Major anthropogenic drivers include groundwater abstraction, land use and land cover change (e.g., urbanization, irrigation, and drainage), and surface-water regulation, which can lower or raise hydraulic heads, modify how precipitation is partitioned between runoff and infiltration (and thus potential recharge), and alter surface-water boundary conditions for aquifers. Sustainable groundwater management aims to keep the cumulative effects of these interventions within bounds that maintain long-term stability of groundwater storage and hydraulic heads across the system. In contrast, unsustainable management leads to persistent groundwater depletion (Gleeson et al. 2020a).

Climate change alters the atmospheric forcing that governs recharge (via changes in precipitation, evapotranspiration, and drought characteristics) and also shifts coastal boundary conditions through sea-level rise and changing storm-, tide-, and wave-driven forcing (IPCC 2021; Smerdon 2017; Richardson et al. 2024). In many dry-summer and semi-arid to arid regions, projected warming and drying increase drought exposure and can suppress diffuse recharge while increasing water demands and groundwater reliance (Smerdon 2017; Spinoni et al. 2021; United Nations 2022). In coastal aquifers with limited recharge, this aridification can further constrain natural recharge and thereby reduce freshwater heads and seaward discharge (Richardson et al. 2024). Groundwater sensitivity to climatic variability across interannual to decadal timescales depends on the climate regime and hydrogeologic response times (hydraulic memory) (Cuthbert et al. 2019a). Aquifers often act as low-pass filters, damping short-term forcing while amplifying multi-annual to decadal groundwater level (GWL) variability (Baulon et al. 2022a). In arid settings, recharge and discharge responses are often weaker and slower, yet hydraulic heads can remain highly sensitive to water-balance shifts (Cuthbert et al. 2019a; Fan et al. 2023a). In contrast, shallow, unconfined aquifers typically respond fastest and provide early indications of change.

When human activities and climate change directly or indirectly lead to groundwater depletion, accessible groundwater storage shrinks and hydraulic heads decline. Wells can run dry, and pumping costs increase as deeper groundwater is more expensive (Perrone and Jasechko 2019). Environmental impacts include reduced groundwater discharge to rivers, springs, and groundwater-fed wetlands, degrading aquatic and riparian ecosystems,

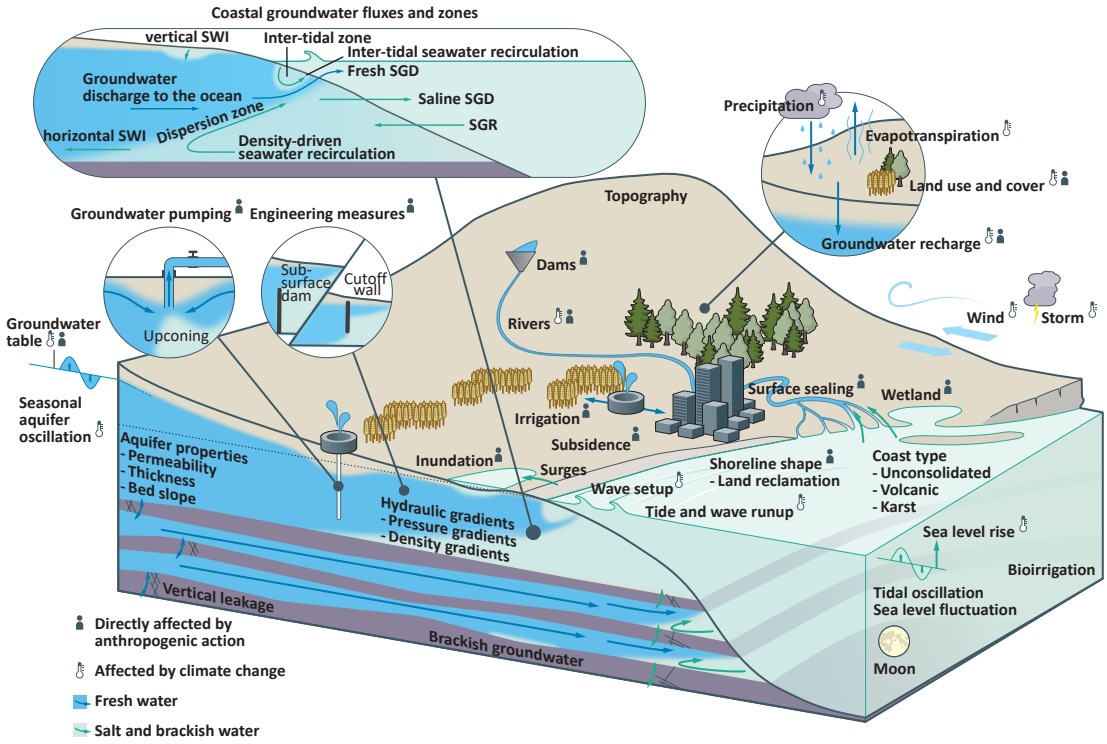


Figure 1: A perceptual model of coastal groundwater flow showing key drivers and limiting factors that shape system dynamics (source: Kretschmer et al. 2025a). Many of these factors are influenced by human activities and climate change and are causally linked. The groundwater hydraulic head integrates the system response, and land–sea hydraulic gradients regulate exchanges with the coastal ocean, including submarine groundwater discharge (SGD) and submarine groundwater recharge (SGR), and the potential for seawater intrusion (SWI). In this thesis, large-sample groundwater level observations provide the observational basis to describe these dynamics and support predictive modeling.

particularly during dry periods (de Graaf et al. 2019; Gleeson et al. 2021), and a net contribution of groundwater depletion to sea-level rise (Konikow 2011). In coastal regions, seawater intrusion (SWI) can occur when land–sea hydraulic gradients are small or landward such as in low-lying coastal plains, allowing saline water to encroach into freshwater aquifers and degrade groundwater quality and availability for drinking, irrigation, industry, and ecosystems (Werner et al. 2013). Documented cases span more than 100 countries and regions worldwide (Cao et al. 2021; Kretschmer et al. 2025a), and risk can increase when inland heads decline (e.g., from pumping or reduced recharge) and when relative sea level rises (Adams et al. 2024; Richardson et al. 2024).

1.2 Large-scale groundwater assessments

While groundwater is exploited and managed locally, its sustainability and impacts are shaped by larger-scale processes and decisions, including its role in the global water cycle and Earth-system functioning, transboundary aquifer governance, globally connected water use (e.g., via trade and virtual water), and international policy agendas (Foster et al. 2013; Guppy et al. 2018; Gleeson et al. 2020a). Yet, much of the existing evidence base is dominated by local and regional case studies, which are difficult to compare and provide limited insight for continental-to-global assessments including model evaluation and prediction (Gleeson et al. 2021). At large spatial scales, groundwater assessments matter in three complementary ways: they (1) provide consistent, comparable evidence on groundwater status and direction of change to help prioritize monitoring and management where risks are emerging; (2) enable comparative analyses that distinguish coherent patterns from site-specific effects and foster more generalizable understanding of controls and Earth-system linkages across hydroclimatic and socio-economic settings; and (3) strengthen inference and prediction by testing diagnostics, assumptions, and model behavior across diverse conditions, thereby revealing robustness and limits to transferability.

Given the fragmented evidence base and the need to derive comparable knowledge about groundwater systems across regions, this thesis focuses on advancing a globally consistent, observation-based evidence base on coastal groundwater dynamics and their temporal evolution. In coastal settings, interacting terrestrial and marine boundary conditions, together with the relevance of groundwater for water security and ecosystems, make such endeavors particularly important. The remainder of this section outlines the observational basis and modeling approaches for large-scale groundwater assessments and describes key challenges.

1.2.1 Observational basis

Groundwater is hidden in the subsurface and therefore difficult to observe directly. Two main types of observations support groundwater assessment: (1) in situ measurements and (2) indirect or proxy observations. In situ measurements provide direct evidence of subsurface conditions through GWLs measured in monitoring wells/piezometers, water-quality and geochemical sampling, pumping records, aquifer-test data (e.g., transmissivity or hydraulic conductivity), and drilling-derived lithologic logs. Field campaigns also sporadically yield recharge estimates inferred from local flux observations (e.g., lysimeters)

or tracer-based approaches, which have been synthesized globally by Moeck et al. (2020). Similarly, quantities such as baseflow-derived groundwater discharge or geophysically inferred subsurface structure are not measured directly in the aquifer but inferred from field observations, and therefore depend on methodological assumptions and interpretation. In contrast, indirect or proxy observations that are not collected as groundwater measurements include satellite remote sensing and standardized meteorological drought indices. These datasets provide broader spatial coverage but only an indirect linkage to subsurface states. Together, these observation types provide evidence on groundwater dynamics and help constrain both process-based and data-driven models.

Among direct in situ observations, GWLs are most widely measured (Alley et al. 2002; United Nations 2022). GWL dynamics reflect the evolving balance of recharge, discharge, and storage. When referenced to land surface (as water-table depth (WTD) in unconfined settings), GWLs provide direct insight into potential accessibility for ecosystems and humans (Reinecke et al. 2024). In coastal settings, GWLs also constrain land–sea hydraulic gradients that govern susceptibility to SWI (Michael et al. 2013; Werner 2017). However, integrating in situ monitoring networks at scale is difficult because monitoring and data stewardship are fragmented across institutions and jurisdictions, digitization and access are limited, and datasets lack standardization regarding formatting, measurement conventions, and metadata. These challenges, together with the importance of comparable, observation-based evidence for groundwater assessments, have been repeatedly highlighted over the past decade (Famiglietti and Rodell 2013; Lall et al. 2020; United Nations 2022; Kuang et al. 2024). Moreover, groundwater data are often scattered, incomplete, and uneven in quality, reflecting persistent limitations in monitoring networks and capacity (Foster et al. 2013). Coastal assessments face the additional constraint that direct monitoring of SWI indicators (e.g., chloride, electrical conductivity) remains sparse (Thorslund and van Vliet 2020; Panthi et al. 2022), while co-located water-quality and water-quantity records are uncommon (Ebeling et al. 2022).

Where in situ observations are sparse, two widely used large-scale alternatives are GRACE (Gravity Recovery and Climate Experiment) satellite gravimetry and standardized meteorological drought indices. GRACE is a satellite-based remote-sensing mission that infers anomalies in terrestrial water storage (TWS) from time-varying changes in Earth's gravity field. Groundwater anomalies are commonly estimated as a residual by subtracting modeled (near-)surface storage changes from GRACE TWS (van Loon et al. 2017). Meteorological drought indices are standardized indicators computed from precipitation alone (SPI) or from a climatic water balance (SPEI) (McKee et al. 1993; Vicente-Serrano et al. 2010). GRACE-based storage anomalies and meteorological drought forcing (SPI/SPEI)

can be computed consistently from gridded datasets and therefore enable large-scale assessments (e.g., Arifin et al. 2025; Babre et al. 2022). However, both approaches face limitations. GRACE-derived groundwater residuals depend on model-based partitioning of TWS and are limited by coarse spatial resolution and land–ocean leakage, which can bias signals in coastal and smaller systems (van Loon et al. 2017; Frappart and Ramillien 2018; Arifin et al. 2025). Meteorological indices represent atmospheric drought forcing, but groundwater responses are filtered by spatially variable lags, nonlinearities, and memory effects, so a single fixed accumulation period can misrepresent groundwater drought conditions across regions (Bloomfield and Marchant 2013; van Loon et al. 2017; Kumar et al. 2016). Accordingly, these products are most useful for large-scale screening and contextualization, and they require careful evaluation against in situ groundwater observations wherever available.

Moving from observation to explanation and prediction at large scales requires spatially consistent information on the main drivers of groundwater dynamics. However, continental-to-global assessments are often constrained by uncertainty in key drivers. Recharge estimates remain model-dependent, i.e. rely on strong assumptions and are only sparsely evaluated against observations in many regions (Reinecke et al. 2021). Groundwater withdrawal estimates are highly uncertain due to sparse reporting and strong spatial and sectoral heterogeneity (Nazari et al. 2025). In addition, globally available datasets that most closely approximate subsurface controls, such as unconsolidated thickness (Zamrsky et al. 2018) or permeability products like GLHYMPS (Huscroft et al. 2018), primarily characterize near-surface materials. They do not resolve three-dimensional aquifer architecture, vertical hydraulic contrasts, or confining units that strongly influence storage, transmissivity, and the timing and magnitude of groundwater responses. As a result, globally consistent driver data needed to explain and predict GWL dynamics, as well as "limiter" data that could constrain groundwater responses in unmonitored regions, remain limited.

1.2.2 Modeling approaches

Because groundwater systems are sparsely observed and highly heterogeneous, large-scale assessments often rely on models to integrate existing observations into spatially and temporally coherent estimates of groundwater states and fluxes, providing information about dominant controls in unmonitored regions and periods and exploration of potential future changes under alternative climate and water-use scenarios (Condon et al. 2021). To be trustworthy, models should be framed as testable hypotheses, with uncertainty made

explicit, and confronted with independent observations not used in model development (Kirchner 2006).

At large scales, groundwater models must balance physical detail with data availability and computational constraints. They can be grouped into three complementary classes: (1) process-based numerical groundwater models that solve governing flow equations (and, where needed, transport equations) for hydraulic head and related fluxes; (2) reduced-complexity or conceptual models that represent dominant groundwater dynamics with simplified structure (including analytical diagnostics); and (3) data-driven models that learn empirical relationships from observations.

Process-based numerical groundwater models such as MODFLOW and FEFLOW are widely used examples of the first model class (Langevin et al. 2017; Diersch 2013). However, process-resolving applications require an explicit, internally consistent representation of the groundwater system, including its key drivers and limiting factors (Figure 1), along with parameterization and evaluation against observed heads and fluxes, often complemented by independent lines of evidence such as baseflow, hydrochemical, or geophysical data (Dams et al. 2012; Rasmussen et al. 2013). The need for such comprehensive system characterization makes these models difficult to scale beyond local to regional scales.

At continental to global scales, process-based groundwater modeling is therefore necessarily simplified and commonly coupled to land-surface or hydrology models that provide recharge and surface-water boundary conditions. For example, Sutanudjaja et al. (2011) demonstrate a large-scale implementation in which MODFLOW is forced by recharge and surface-water boundary conditions from a global hydrological/land-surface model built from global datasets. Another approach is to represent lateral groundwater flow directly within global hydrological models using simplified process formulations (Reinecke et al. 2019). Finally, reduced-complexity process-based hydrological models can be used to infer key groundwater-relevant quantities such as recharge from more widely available meteorological and streamflow data (Nolte et al. 2021). Reduced-complexity approaches also include analytical frameworks that, for example, provide scalable diagnostics of groundwater response times and sensitivity to climate forcing (Cuthbert et al. 2019a). Across these approaches, uncertainty in key drivers, coarse parameter fields, and sparse independent head and flux observations for evaluation jointly increase structural and parameter uncertainty and reduce confidence in attributing simulated changes to specific drivers or processes. Large-scale simulations should therefore be interpreted as spatially complete, testable representations rather than definitive reconstructions (Condon et al. 2021).

In parallel, data-driven models based on machine learning (ML), particularly deep learning (DL), are increasingly used to model GWLs by learning empirical relationships from observed time series and predictor data (Rajaee et al. 2019; Tao et al. 2022). However, most groundwater applications remain single-site and depend on carefully curated, high-quality GWL time series (Chidepudi et al. 2023; Wunsch et al. 2022). In surface hydrology, standardized large-sample datasets with catchment attributes (e.g., CAMELS; Addor et al. 2017) have enabled rainfall–runoff (streamflow) modeling across hundreds of U.S. catchments using a single multi-site DL model (Kratzert et al. 2019). In groundwater predictive modeling, scaling is more difficult because sites cannot be linked to an equally consistent spatial unit and descriptor set, given sparse and heterogeneous subsurface characterization (Heudorfer et al. 2024; Chidepudi et al. 2025).

Large-scale data-driven approaches either scale across observation networks, learning across many monitored wells but remaining limited to locations with GWL time series and consistent driver/limiter covariates (Heudorfer et al. 2024; Chidepudi et al. 2025), or pursue spatial completeness through regionalization (Hrachowitz et al. 2013). Both strategies rely on generalization, but in different ways. Multi-site modeling requires generalization across monitored wells to enable information sharing within observation networks (Kratzert et al. 2019), whereas regionalization requires generalization beyond the monitored sample to less monitored or unmonitored locations. This can involve transfer of groundwater dynamics from well-monitored sites to target sites with sparse or missing GWL records (e.g., Haaf et al. 2023), or spatial generalization beyond monitored locations by learning relationships between groundwater observations and spatially continuous covariates to predict groundwater variables across the study region (e.g., WTD across the United States in Ma et al. 2024). As with large-scale process-based modeling, however, spatially continuous data-driven modeling is constrained by uncertainty in gridded inputs and by limited transferability of learned relationships across contrasting hydrogeologic settings. Moreover, because purely data-driven models can reproduce patterns without representing mechanisms, they do not, by themselves, enable causal attribution, motivating hybrid and explainable approaches that improve interpretability and help build confidence in ML-based predictions for scientific inference and, where appropriate, decision support (Reichstein et al. 2019; Shen and Lawson 2021).

1.2.3 Knowledge trade-off and strategies

Taken together, the observational constraints and the limitations of large-scale models outlined above imply a fundamental trade-off in groundwater assessment: spatially

complete information is typically obtained only at the cost of stronger assumptions and reduced local fidelity, whereas locally faithful information remains geographically uneven and difficult to compare.

Large-scale inference in groundwater research therefore proceeds along two complementary strategies: (1) top-down approaches that yield spatially continuous but assumption-dependent estimates from models and aggregated signals; and (2) bottom-up approaches that standardize and synthesize large samples of in situ groundwater observations (specifically GWL time series) to retain local fidelity despite heterogeneous coverage. The latter parallels the "large-sample hydrology" perspective developed in surface-water research, which emphasizes learning generalizable patterns from standardized analyses across many sites while retaining local realism (Gupta et al. 2014). This thesis adopts the bottom-up strategy, a promising but comparatively underexplored direction in groundwater research.

1.3 Research gaps and directions

Building on the observational basis and modeling strategies outlined above, key gaps persist in the availability and standardized use of groundwater observations for comparable large-scale assessments of coastal groundwater dynamics. These gaps motivate the specific research directions pursued in this thesis.

Foremost, a harmonized, globally comparable set of in situ GWL time series is needed to enable such assessments, yet these data remain sparse and fragmented. Analogous to large-sample hydrology (Addor et al. 2020; Gupta et al. 2014), initiatives are emerging to fill this gap, including a vision for a global groundwater platform for groundwater monitoring and prediction (Condon et al. 2021) and the Global Groundwater Monitoring Network (GGMN), which now offers a centralized portal with open access to directly downloadable GWL time series (IGRAC 2025). Despite these advances, no open, research-ready global compilation of GWL time series was available as of early 2025. The most widely used global reference remains the steady-state WTD compilation by Fan et al. (2013), which underpins many groundwater studies (e.g., Shokri–Kuehni et al. 2020). However, because it lacks temporal information, it cannot inform about groundwater dynamics.

Second, there is a critical need to understand groundwater fluxes across boundaries and related Earth-system linkages, including groundwater discharge and exchange across the land–sea boundary (Blöschl et al. 2019). In coastal aquifers, these exchanges are regulated by hydraulic heads (including their position relative to sea level) and the resulting land–sea

hydraulic gradients (Kretschmer et al. 2025a; Jasechko et al. 2020). Yet coastal-specific, observation-based evidence on the status and change of heads and gradients remains limited at large scales, constraining the ability to anticipate risks to water security and ecosystems and to prioritize monitoring and management as conditions evolve. Recent continental-to-global syntheses of in situ GWL time series provide important evidence on groundwater change (Chávez García Silva et al. 2024; Jasechko et al. 2024), but coastal aquifers remain underrepresented and have not been assessed at the global scale. As a result, it remains poorly quantified where coastal land–sea gradients are inland- versus seaward-directed, and where head trends are shifting them over time, limiting understanding of land–sea exchange at the global scale and the identification of coastal regions where aquifers may be particularly susceptible to SWI.

Third, despite pronounced local heterogeneity, it remains unclear to what extent coastal groundwater dynamics exhibit globally recurring patterns that enable systematic comparison. Local and regional studies show that groundwater hydrographs and their drivers can vary markedly over short distances (Wunsch et al. 2022), highlighting the challenge of distinguishing site-specific variability from functional similarity – comparable groundwater responses under comparable controls – across scales (Blöschl et al. 2019). Hydrograph-based typologies that summarize time series into diagnostic signatures provide a standardized description of groundwater dynamics (Heudorfer et al. 2019). They can reveal coherent dynamic regimes at large scales and offer diagnostic benchmarks for large-sample model evaluation by testing whether models reproduce these regimes across comparable settings (Gupta et al. 2014; Barthel et al. 2021). However, it remains largely unquantified whether and how typology-derived regimes recur at the global scale and whether globally consistent descriptor datasets, which largely represent surface conditions and provide only coarse proxies for subsurface structure and key drivers (Figure 1), can explain observed coastal groundwater dynamics across heterogeneous groundwater systems worldwide.

Fourth, predictions at large scales that span different hydroclimates and hydrogeological settings, and that remain useful under typical monitoring constraints (short, gappy, irregular, and noisy records) are needed. Both process-based and data-driven models struggle with predictive modeling beyond monitored locations (a top-down goal), where limited observations and heterogeneous aquifer properties restrict regionalization (Hrachowitz et al. 2013). However, compared to often infeasible process-based, site- and process-resolving approaches at large scales, DL enables joint learning and predictive modeling across many monitoring locations. Bottom-up modeling is feasible when large samples of harmonized GWL time series are available together with static site descriptors

that allow the model to distinguish wells and represent site-specific effects and, where informative and spatially continuous, can support spatial generalization across similar settings (Kratzert et al. 2019; Kratzert et al. 2024; Heudorfer et al. 2024). Yet for groundwater, evidence on scalable multi-site DL remains limited, and key questions remain about robustness under typical GWL data constraints and about generalization across wells, given also limited and uncertain static site descriptors (Heudorfer et al. 2024; Chidepudi et al. 2025), especially at the global scale. Systematic benchmarking and targeted model development are therefore needed to assess the feasibility, benefits, and limits of scalable DL approaches for GWL prediction under real-world data constraints.

1.4 Thesis outline and research questions

These research gaps motivated three overarching research questions (RQs) addressed in this thesis. Chapter 2 introduces the Coastal GroundWater Level (CGWL) dataset as a core thesis output and the global data baseline for all subsequent analyses. Chapters 3–5 constitute the main research chapters and address RQ1–RQ3, respectively. Chapter 6 concludes with a discussion of implications, limitations, and outlook.

RQ1: How is land–sea exchange and susceptibility to seawater intrusion characterized along global coasts?
This question responds to limited coastal-specific, observation-based evidence at the global scale on the status and change of groundwater heads and land–sea hydraulic gradients, which regulate land–sea exchange and shape SWI susceptibility. Specifically, it addresses where present-day conditions suggest elevated susceptibility to SWI and where these conditions may be changing as GWLs rise or fall. Chapter 3 addresses this by deriving two complementary indicators from in situ GWL observations: land–sea hydraulic gradients as an indicator of SWI-relevant conditions, and recent GWL trends as an indicator of how these conditions are evolving. Together, these indicators enable a spatially comparable global groundwater assessment of conditions that regulate land–sea exchange and help identify coastal regions where head changes are shifting gradients over time and, by implication, potential susceptibility to SWI.

RQ2: What types of recurring regimes of coastal groundwater dynamics emerge globally, and to what extent can they be explained?
This question responds to the knowledge gap that, despite pronounced local hydrogeological heterogeneity, it remains unclear whether coastal groundwater dynamics exhibit robust, recurring regimes that enable systematic comparison at the global scale, and

whether globally consistent (largely surface-based) datasets capture the dominant controls of these regimes. Chapter 4 addresses this by deriving a hydrograph-signature typology of coastal groundwater dynamics from globally distributed GWL time series, investigating how the inferred regimes persist across analysis scales, and using explainable ML to relate regimes to globally available descriptors (i.e., proxies for key drivers and limiters; Figure 1). This provides a data-driven basis to identify the controls that most consistently differentiate regimes across regions and to make remaining explanatory gaps, and thus limits to large-scale generalization, explicit.

RQ3: How can coastal groundwater levels be predicted at the global scale with deep learning under realistic data limitations?

This question targets the problem that scalable multi-site groundwater prediction is rarely benchmarked under typical monitoring constraints, and that the static site descriptors commonly used to support spatial generalization are often missing, inconsistent, or uncertain in hydrogeology. Chapter 5 responds by applying multi-site DL model variants across diverse coastal regions using a global dataset, and by quantifying how predictive skill depends on data availability in GWL time series and the scope of training. It further presents trainable site representations (1) as a scalable alternative when static descriptors are unavailable or unreliable; and (2) as a basis for post hoc analyses of the model's internal similarity structure and behavior across wells, helping to assess current limits to generalization in DL-based predictive modeling of groundwater dynamics.

2 The Coastal GroundWater Level (CGWL) dataset

The Coastal GroundWater Level (CGWL) dataset is a global compilation of in situ GWL observations acquired between 2019 and 2025 from various governmental agencies and data portals that are listed in Table 1. The published dataset version includes records cleared for publication and is released under a CC-BY-NC license. This chapter provides an overview of the dataset and comparison with the global observational water table depth (WTD) compilation by Fan et al. (2013). Detailed documentation of data sources, preprocessing and harmonization, dataset structure, and spatiotemporal coverage is provided in Appendix A.

The data preprocessing involved applying consistent criteria for data selection, quality control, and metadata harmonization (Section A.1). The final dataset comprises approximately 629,000 well records (480,000 unique locations at 3-decimal coordinates) from WTDs shallower than 100 m and wells within 100 km of the coastline. This inland extent reflects evidence that SWI and coastal groundwater impacts can extend several tens of kilometers inland in low-lying or hydrologically stressed aquifers (Barlow and Reichard 2010; Mueller et al. 2024). It also aligns with broader coastal-zone definitions in groundwater and water resource research that often extend inland to account for catchment-scale processes (Bosserelle et al. 2022), ecological and economic linkages (Martínez et al. 2007), and urban zones (Cao et al. 2021).

In the course of the dataset development, land-surface elevation data at the well location were required for two steps: (1) to harmonize GWLs reported relative to mean sea level to a consistent ground-surface reference (WTD), and (2) to compute groundwater elevation as surface elevation minus WTD. Where elevations were missing in the source dataset (21% of wells), they were taken from CoastalDEM (Kulp and Strauss 2019) or SRTM (NASA JPL 2020). The residual uncertainty introduced by DEM-based elevations and reference-point harmonization is estimated at ~1-3 m (Appendix A.1).

Hydrogeologically meaningful aggregation and aquifer differentiation are currently limited at large scales (Appendix A.2). Aquifer state (e.g., confined or unconfined) and other subsurface attributes are rarely reported together with GWL observations, and available global layers cannot reliably identify which aquifer (in multi-layer systems) or hydrogeologically coherent unit individual wells tap, nor delineate the contributing area for surface controls (an inherently difficult task even with dense observations and models).

Table 1: List (region-wise alphabetical) of references from which GWL time series were obtained. Checkmarks indicate inclusion in the analyses presented in Chapters (Ch.) 3–5.

Number	Source (government authority / portal)	Region	Access year	Ch. 3	Ch. 4	Ch. 5
1	Australian Groundwater Explorer – Bureau of Meteorology (Australia)	Australia	2021	✓	✓	✓
2	DOV (Datenbank Ondergrond Vlaanderen) – Vlaamse overheid (Belgium – Vlaanderen)	Belgium	2020	✓	✓	✓
3	RIMAS (Rede Integrada de Monitoramento das Águas Subterrâneas) – Serviço Geológico do Brasil (Brazil)	Brazil	2022	✓	✓	
4	Department of Environment and Climate Change (Canada — Nova Scotia)	Canada	2021	✓	✓	✓
5	GIN (Groundwater Information Network) — Natural Resources Canada (Canada)	Canada	2021	✓	✓	✓
6	Ministry of Environment and Climate Change Strategy (Canada -– British Columbia)	Canada	2021	✓	✓	✓
7	Nationale Geologiske Undersøgelser for Danmark og Grønland (GEUS; Denmark)	Denmark	2019	✓	✓	✓
8	Waterbase — European Environment Agency (EEA; European Union)	Europe	2021	✓		
9	ADES (Portail national d'accès aux données sur les eaux souterraines) -– Bureau de Recherches Géologiques et Minières (BRGM; France)	France	2020	✓	✓	✓
10	Landesamt für Umwelt (LfU; Germany -– Schleswig-Holstein)	Germany	2020	✓	✓	✓
11	Niedersächsischer Landesbetrieb für Wasserwirtschaft, Küsten- und Naturschutz (NLWKN; Germany -– Niedersachsen)	Germany	2021	✓	✓	✓
12	Staatliches Amt für Landwirtschaft und Umwelt Mittleres Mecklenburg (Germany — Mecklenburg-Vorpommern)	Germany	2020	✓	✓	✓
13	Staatliches Amt für Landwirtschaft und Umwelt Vorpommern (Germany – Mecklenburg-Vorpommern)	Germany	2021	✓	✓	✓
14	GGMN (Global Groundwater Monitoring Network) dataset and Jasechko et al. (2024) annual groundwater level dataset – International Groundwater Resources Assessment Centre (IGRAC)	Global	2025	✓		
15	Instituto Geológico y Minero de España (IGME; Spain)	Spain	2020	✓		
16	India Water Resource Information System (India-WRIS) — Ministry of Jal Shakti (India)	India	2020	✓		✓
17	Environmental Protection Agency (EPA; Ireland)	Ireland	2020	✓	✓	
18	Hydstra -– Department of Water and Sanitation (DWS; South Africa)	South Africa	2021	✓	✓	✓
19	NGA -– Department of Water and Sanitation (DWS; South Africa)	South Africa	2019	✓	✓	✓
20	Sveriges Geologiska Undersökning (SGU; Sweden)	Sweden	2021	✓	✓	✓
21	DINOloket -– Geologische Dienst Nederland (TNO; Netherlands)	The Netherlands	2020	✓	✓	✓
22	California Department of Water Resources (United States of America — California)	United States of America	2022	✓	✓	✓
23	NWIS (National Water Information System) — United States Geological Survey (USGS; United States of America)	United States of America	2020	✓	✓	✓

To enable area-representative summaries and cross-regional comparability despite uneven monitoring density, GWL observations were additionally aggregated into near-equal-area global grid cells (H3 hexagons at resolutions 8 (H3-8) and 6 (H3-6); mean cell areas ∼0.74 and ∼36 km^2; Uber Technologies Inc 2024) and into coastal surface-water catchments (HydroBASINS level 12; mean basin area ∼141 km^2; Lehner and Grill 2013) (Section A.3). While aggregation cannot create uniform global monitoring coverage or resolve aquifer-specific signals, it makes the uneven well distribution more comparable across regions and supports large-scale analyses.

The CGWL dataset includes GWL time series, summary statistics, well attributes, and spatially aggregated products of groundwater depth and elevation, all linked via identifiers (Section A.4). Together, these components provide a consistent framework to examine spatiotemporal patterns in coastal groundwater and to compare conditions across regions and analysis scales. The dataset includes observations spanning more than a century and covers coastal regions worldwide, representing the most extensive and diverse compilation of in situ coastal GWL observations currently available. However, observations remain unevenly distributed in space and time, with the densest coverage in North America, northwestern Europe, and Australia (Figure 2; Section A.5). High-resolution and complete time series are not uniformly available across regions, and many wells have only sparse or single measurements. These spatiotemporal characteristics of the CGWL dataset, together with limited aquifer differentiation (Section A.2) and less quantifiable biases, such as underrepresentation of dry or decommissioned wells (Hora et al. 2019), should be considered when interpreting global representativeness.

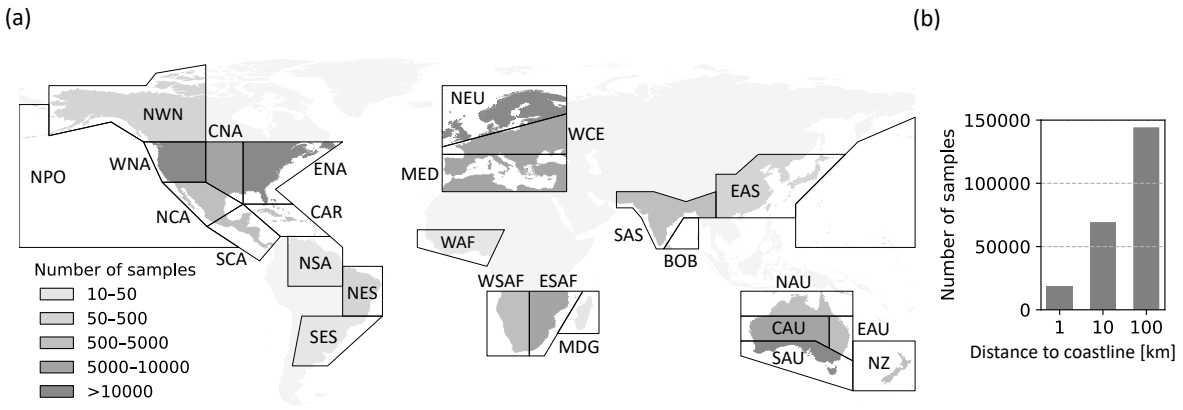


Figure 2: Spatial coverage of the CGWL dataset. (a) Spatial distribution of observations after aggregation to H3 hexagons (resolution 8; Uber Technologies Inc 2024), shown within IPCC-defined subregions (Iturbide et al. 2020). (b) Number of observations by distance band from the coastline.

Steady-state aggregated WTD values from the CGWL dataset were compared with the

only existing global observational WTD dataset by Fan et al. (2013). Within the 100 km coastal ribbon, the CGWL dataset contains roughly twice the observational support of the Fan et al. observational dataset (OBS) (~252,000 wells). Both compilations heavily sample coastal regions in North America, Western Europe, and Australia, but the CGWL dataset provides notably denser coverage in India, South Africa, and Mexico, and spans a more diverse set of northwestern European coasts, whereas Fan et al. OBS includes comparatively more sites along the Brazilian coast. For a consistent spatial comparison, Fan et al. OBS data and simulated equilibrium water tables (Fan et al. 2023b) were aggregated to the H3-8 grid. When restricting the comparison to grid cells with overlapping coverage (Figure 3a), the CGWL WTD distribution closely matches that of Fan et al. OBS. Both datasets show a near-identical proportion of shallow ($\leq$5 m) and deeper (>5 m) WTDs for the same areas (50% shallow in CGWL vs. 51% in Fan et al. OBS; respective median WTDs: 4.9 m vs. 4.8 m). In contrast, the simulated estimates show stronger clustering of shallow water tables and a larger share of deeper WTDs (>15 m), while underrepresenting mid-range depths (2–15 m). Details on model uncertainty and process representation are provided in Fan et al. (2013) and Reinecke et al. (2024). Despite the overall agreement between the CGWL dataset and Fan et al. OBS across shared grid cells (Pearson $r = 0.95$), substantial local differences remain: 17% of grid cells differ by more than 1 m (Figure 3b), and among grid cells with shallow WTDs ($\leq$5 m in either dataset), this fraction is about 12%.

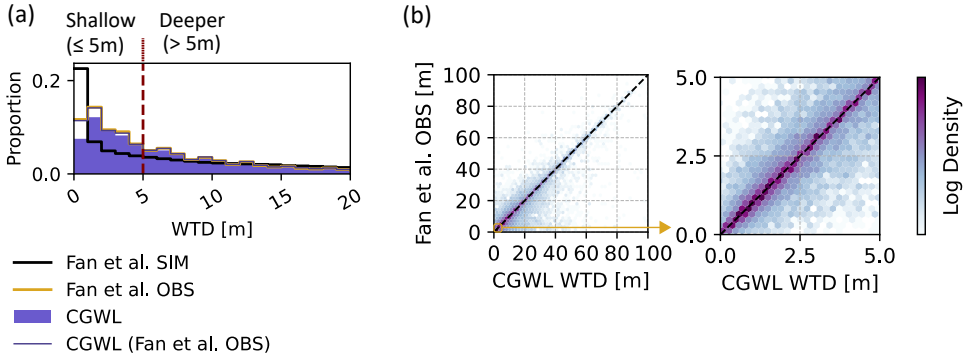


Figure 3: Comparison of water table depths (WTDs) in the CGWL dataset with previous global WTD estimates. (a) WTD distributions from the CGWL dataset, Fan et al. (2013) observations (OBS), and the Fan et al. (2023b) simulation (SIM). The vertical dashed line marks a 5 m threshold between shallow and deeper water tables. (b) Density plot comparing CGWL WTD (x-axis) and Fan et al. (OBS) WTD (y-axis) (Pearson $r = 0.95$). The color scale indicates point density on a logarithmic scale, and the inset on the right shows the shallow range (0–5 m).

A key added value of the CGWL dataset beyond enhanced coastal coverage is that it provides measurement timestamps and, where available, time series, enabling spatiotemporal analyses. The analyses in Chapters 3–5 were conducted using temporal snapshots

and study-specific subsets of the CGWL dataset, reflecting data availability at the time of each analysis and inclusion criteria aligned with each research objective (Table 1).

3 Coastal groundwater level trends reveal global susceptibility to seawater intrusion

3.1 Abstract

Coastal groundwater is a vital freshwater source threatened by overabstraction and sea-level rise, yet global patterns of declining GWLs and susceptibility to SWI remain poorly constrained. Here, we present the first global assessment based on in situ observations from ~ 480,000 coastal monitoring locations. From 1990 to 2024, 21% of grid-cell aggregates had significant rising or falling GWL trends with magnitudes $\geq 0.1\,\mathrm{m\,yr^{-1}}$ and with declines becoming more frequent in the last 9 years. More pronounced changes are observed for deeper water tables ($\rho_s = 0.63$), in arid settings ($\rho_s = 0.56$), and in some rural areas. SWI susceptibility is higher where seaward freshwater discharge is weak or gradients reverse landward, leaving limited hydraulic resistance to SWI. Decadal trend extrapolation suggests that these conditions mostly persist (93.4%), while 3.5% newly emerge and 3.1% stabilize (gradients strengthen seaward). The results provide global evidence for prioritizing monitoring and management of coastal groundwater at risk of salinization.

3.2 Introduction

Groundwater is a critical freshwater source in coastal regions, which are home to 2.86 billion people (Cosby et al. 2024), and plays a vital role in sustaining coastal ecosystems and their services (Dyring et al. 2022; Luijendijk et al. 2020) as well as marine biota (Lecher and Mackey 2018). However, coastal aquifers face pressure from growing coastal populations (Neumann et al. 2015), and are increasingly vulnerable to SWI, which threatens both the quantity and quality of available water. Salinization reduces water suitability for human use. It has adverse health impacts when drinking-water salinity increases (Mueller et al. 2024), threatens agriculture by lowering soil fertility and crop yields when salinity of irrigation water increases (Ghirardelli et al. 2024), and affects coastal ecosystems, where increased salinity can stress vegetation and soil communities and lead to shifts in species composition (Venâncio et al. 2022; Tully et al. 2019). Climate change and, as a consequence, sea-level rise, changes in groundwater recharge, and intensifying water use increase pressure on coastal groundwater systems (IPCC 2021; Spinoni et al. 2021; Richardson et al. 2024). This makes it increasingly important to identify where coastal groundwater resources are susceptible to SWI and where monitoring

and management should focus to detect and respond to emerging threats.

In coastal settings, hydraulic gradients at the land–sea boundary govern the balance between fresh submarine groundwater discharge and the intrusion of seawater into the nearshore aquifer. Persistent landward or near-zero seaward gradients therefore favor lateral SWI and constitute a first-order control on coastal groundwater–ocean exchange (Werner et al. 2012; Werner 2017; Kretschmer et al. 2025a). Here, we use the hydraulic gradient as a scalable proxy for whether freshwater tends to discharge seaward or seawater tends to intrude landward. Gradients summarize coastal head differences and thus integrate the effects of multiple controls. Aquifer properties and architecture, including surface-water connectivity, modulate how natural recharge, human pressures (e.g., groundwater abstraction, land drainage, land-surface changes that modify recharge), and sea-level rise translate into gradients. In particular, two natural settings predispose coasts to landward or near-zero seaward gradients by constraining freshwater heads: topography-limited low-relief coastal plains and recharge-limited (often arid and water-limited) settings (Michael et al. 2013). In topography-limited settings, freshwater heads are already close to the land surface, leaving limited capacity for further head increase under sea-level rise. In recharge-limited settings, limited replenishment suppresses freshwater heads and constrains head recovery. Here, long groundwater response times (hydraulic memory) can delay and buffer responses to climate variability, even as water table sensitivity to recharge or precipitation can be high and increase with aridity (Fan et al. 2023a; Cuthbert et al. 2019a). Projected aridification and shifting precipitation regimes are expected to further reduce recharge in recharge-limited coasts, increasing SWI susceptibility (Richardson et al. 2024).

Yet despite our general process understanding, the global spatiotemporal distribution of SWI-susceptible conditions remains poorly characterized. This limits our ability to anticipate where and how salinization pressures may intensify under future climatic and anthropogenic change and reflects broader limitations in global groundwater assessments, which often lack the resolution, integration, and observational grounding needed to capture the complexity of groundwater systems (Gleeson et al. 2020a). Most assessments of coastal groundwater systems are focused on local or regional case studies (e.g., Morgan and Werner 2015; Costall et al. 2020), which presents challenges of commensurability when seeking a better understanding about global-scale processes (Gleeson et al. 2021). Global approaches to model SWI (Adams et al. 2024; Zamrsky et al. 2024; Kretschmer et al. 2025b) continue to face high uncertainties due to assumptions concerning boundary conditions and hydrogeological parameterization, and are challenging to evaluate due to the coarse spatial resolution and spatially-biased availability of time-varying GWL

data (Gleeson et al. 2021; Reinecke et al. 2024). GRACE (Gravity Recovery and Climate Experiment) satellite observations provide valuable information for major groundwater basins. However, coarse spatial resolution, land–ocean leakage near coasts, and the residual-based derivation of groundwater storage changes from model-partitioned water components can yield physically implausible estimates and limit applicability, particularly for smaller or coastal systems (Frappart and Ramillien 2018; Costantini et al. 2023; Arifin et al. 2025).

In contrast, key indicators of coastal groundwater susceptibility to SWI can be derived from in situ GWL observations, specifically hydraulic heads that favor lateral SWI and declining GWLs that lower freshwater heads (Kretschmer et al. 2025b; Ferguson and Gleeson 2012; Jasechko et al. 2020). However, GWLs remain difficult to integrate at scale due to fragmented responsibilities, limited digitization, accessibility issues, and non-standardized datasets (United Nations 2022; Famiglietti and Rodell 2013; Lall et al. 2020). Despite these challenges, few efforts have leveraged large-sample in situ GWL data. Fan et al. (2013) were the first to compile GWL observations globally. This was the basis for developing modeled estimates of water table depth (WTD) with global coverage (e.g., Ma et al. 2024; Fan et al. 2013), now widely used in groundwater-related environmental studies (e.g., Shokri–Kuehni et al. 2020). Recent in situ syntheses have quantified GWL trends across terrestrial aquifer systems globally (Jasechko et al. 2024) and regionally in southwestern Europe (Chávez García Silva et al. 2024), but without a specific coastal focus. By contrast, Jasechko et al. (2020) focused on coastal areas but limited their analysis to the contiguous United States, while Nolte et al. (2024) used thousands of globally distributed coastal GWL time series to identify common GWL dynamics and highlight limitations of global datasets in explaining these patterns, without deriving GWL trends or explicitly assessing susceptibility to SWI. Collectively, these studies underscore both the value of in situ GWL data and the persistent gap in globally integrated, observation-based assessments of coastal aquifers. Here, we compile an unprecedentedly large and diverse global set of in situ coastal GWL observations to derive land–sea hydraulic gradients and recent trends summarized on a common spatial grid for global mapping as observation-based indicators of SWI susceptibility, and to examine how these indicators vary across aridity conditions and coastal regions worldwide.

3.3 Susceptible hydroclimatic conditions

Hydroclimatic susceptibility to lateral SWI in our framework is defined by two variables: the land–sea hydraulic gradient that we use as an observation-accessible screening di-

agnostic, and aridity as a practical proxy for natural recharge limitation. We quantify the land–sea gradient as the change in hydraulic head between inland monitoring locations and the coastline, divided by their distance to the coastline, and classify it as "flat" when it is landward or weakly seaward ($\leq 10^{-3}\,\mathrm{m\,m^{-1}}$). This "flat-gradient" class favors lateral SWI (Kretschmer et al. 2025b; Ferguson and Gleeson 2012). All remaining gradients are grouped as steep ($> 10^{-3}\,\mathrm{m\,m^{-1}}$), representing stronger seaward head gradients that generally oppose inland SWI. While this steep class spans a wide range of seaward conditions, we use this binary split as a simple, transparent global screening. Aridity is represented by the aridity index (AI; the ratio of precipitation to potential evapotranspiration; Zomer et al. 2022), with water-limited climates (AI$\leq$1; rather dry) contrasted against energy-limited climates (AI>1; rather wet). The combination of flat gradients and water-limited climates forms the most susceptible hydroclimatic cluster (C1) in our analysis (Figure 4; compare Methods). WTD provides important context because shallow groundwater implies limited vadose storage, so modest declines can more readily weaken land–sea gradients and increase salinization risk. Globally, aquifers near coasts are often shallow: the median coastal WTD is 6.1 m, and nearly half of all observations are shallower than 5 m (Chapter 2). Flat gradients, which weaken the natural flow of groundwater toward the sea, are found in about one-third of all observed coastal areas. Although steeper gradients dominate overall (skewed distribution in Figure 4a), flat gradients remain common even tens of kilometers inland (34-37% of all observations; Figure 4c). Compared to the global aridity distribution of coastal regions, our dataset shows a distinct clustering near the transitional AI value of approximately 1 (marking the boundary between water- and energy-limited climates), a zone in which groundwater recharge is expected to be particularly sensitive to modest shifts in long-term aridity (Berghuijs et al. 2024). In contrast, moderately arid regions (AI~ 0.3-0.7) and humid regions (AI»1) are underrepresented in our dataset (Figure 4b). Overall, 74% of observations fall within water-limited conditions (AI$\leq$1). Importantly, flat-gradient, water-limited conditions (C1; most susceptible) are most frequent within 1 km of the coastline, where they account for 28% of observations, but they also occur farther inland (Figure 4c). In Figure 4d, C1 forms contiguous coastal belts in regions such as the southeastern U.S. (ENA), Gulf of Mexico (CNA and SCA), and northeast Australia (parts of CAU and NAU). Steep-gradient, water-limited conditions (C2) dominate across all coastal distances (Figure 4c). Near the coast, C2 accounts for 52% of observations. By contrast, susceptibility conditions, where gradients are flat but recharge is not limiting (C3), are least common.

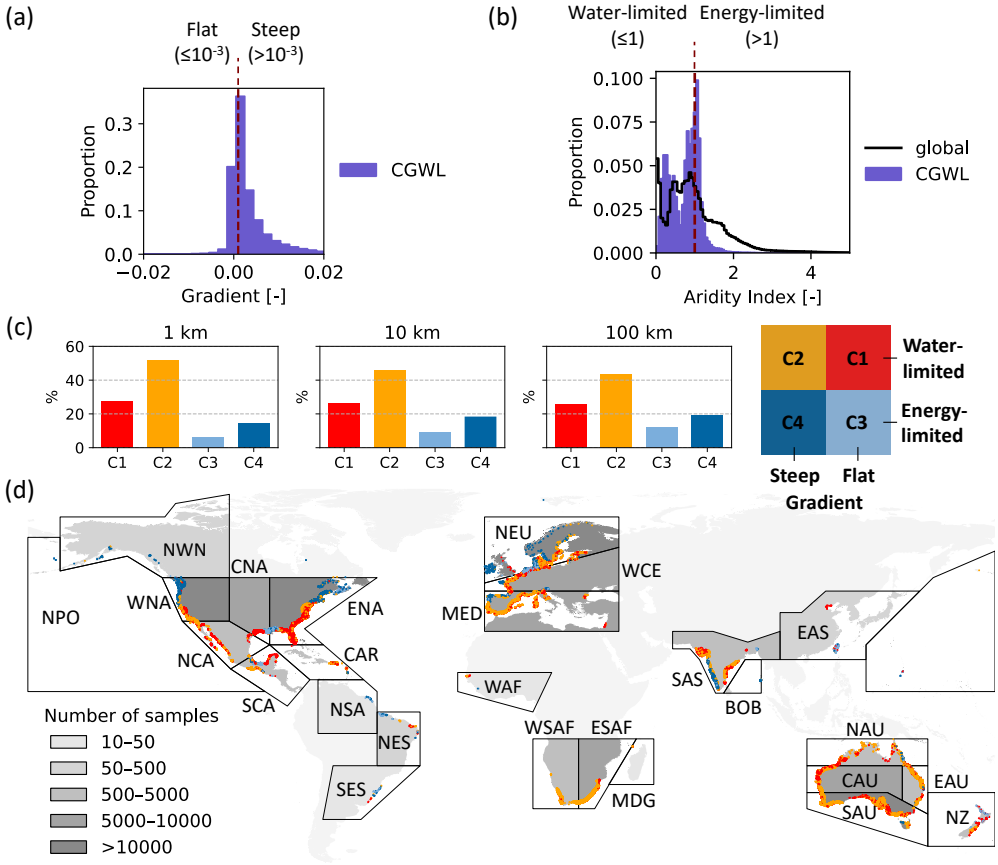


Figure 4: Global patterns of hydroclimatic conditions indicating susceptibility of coastal groundwater systems to seawater intrusion. (a) Distribution of land–sea hydraulic gradients derived from the CGWL dataset, with a red dashed line marking the threshold separating "flat" ($\leq 10^{-3}$ m m^{-1}; landward or near-zero seaward gradients) from "steep" seaward gradients. (b) Distribution of aridity index (AI) values (Zomer et al. 2022) for the CGWL dataset and coastal regions globally, with a red dashed line marking the climatic threshold (AI~ 1) used to distinguish water- and energy-limited regimes. (c) Proportions of four hydroclimatic susceptibility clusters (C1–C4) at three distances from the coastline (1 km, 10 km, 100 km). (d) Global distribution of categorized observations (plotted in grid cell centers with random overlap of clusters), shown together with the number of samples in IPCC reference regions (Iturbide et al. 2020) with ≥10 observations.

3.4 Prevalence of trends and sensitivity to record length

Changes in GWLs alter freshwater heads and, therefore, the land–sea hydraulic gradients. For the 34-year period 1990–2024 studied, our trend analyses show that 28% of observations exhibit statistically significant GWL changes exceeding ± 0.1 m yr^{-1} over 9-year windows. 21% of observations exhibit such trends when evaluated over 19-year windows. The frequency of upward and downward changes in GWLs is balanced across both periods,

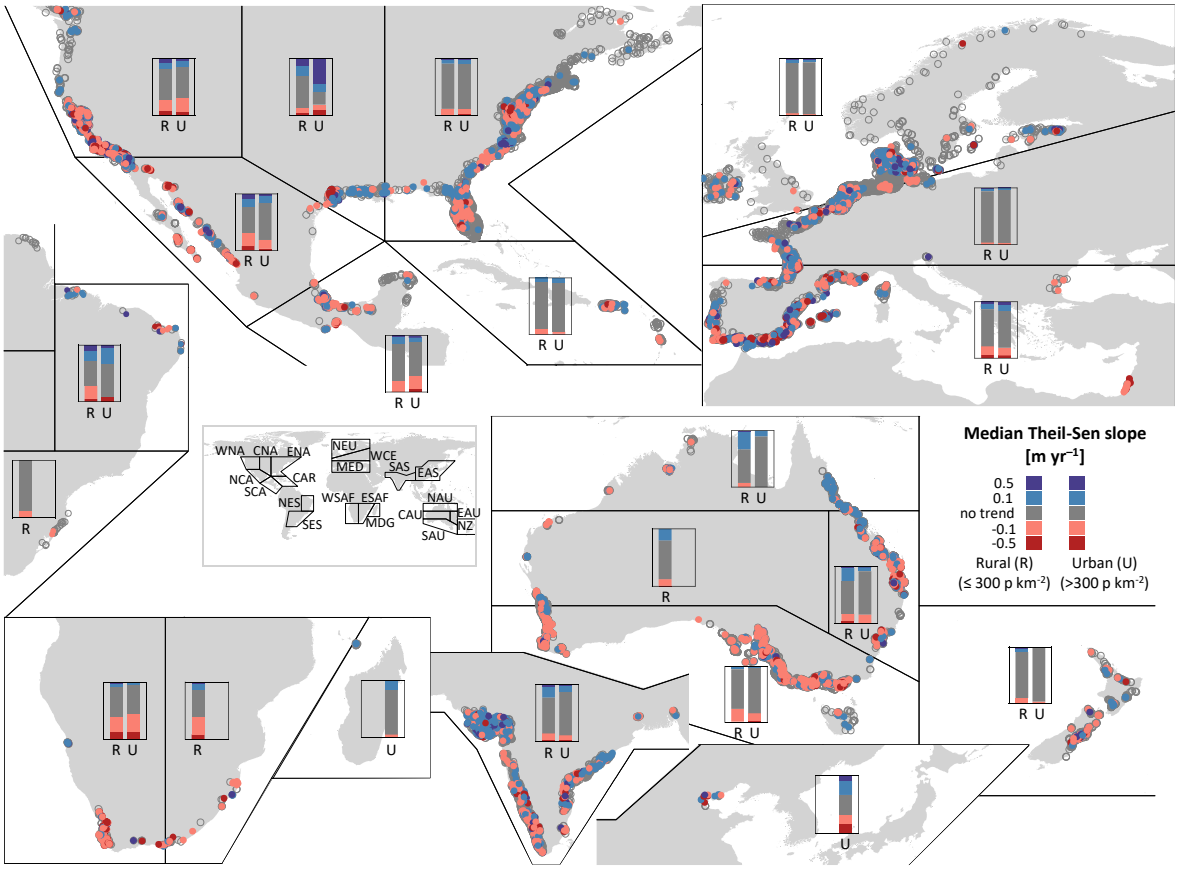


Figure 5: Spatial patterns of recent GWL trends across global coastal zones, grouped by IPCC regions (Iturbide et al. 2020). Colored points on maps represent individual observations that are plotted in grid cell centers with random overlap of Theil-Sen slope trend categories on top of "no trend": strong decline ($< -0.5\,\mathrm{m\,yr^{-1}}$, dark red), moderate decline (-0.5 to $-0.1\,\mathrm{m\,yr^{-1}}$, light red), no trend ($-0.1$ to $0.1\,\mathrm{m\,yr^{-1}}$ and/or non-significant, gray), moderate rise (0.1 to $0.5\,\mathrm{m\,yr^{-1}}$, light blue), and strong rise ($> 0.5\,\mathrm{m\,yr^{-1}}$, dark blue). Bar charts show the proportions of trend categories per region ($\geq$10 observations), stratified by rural ($\leq$300 people $\mathrm{km^{-2}}$) and urban (>300 people $\mathrm{km^{-2}}$) areas (CIESIN 2018). Supporting regional trend results are provided in Appendix B.5.

with a slight bias toward declines in the shorter timeframe (42% upward compared to 58% downward among absolute trends). To evaluate how strongly record length affects inferred trends, we compared slopes from overlapping 9- and 19-year windows for the subset of observations with both time frames available (Appendix B.2; B.3). These are moderately correlated across windows (Spearman $\rho_s = 0.58$). Trend direction is consistent in most observations (76%), while ~ 19% are ambiguous. Here, the 9- and 19-year slopes have opposite signs, but at least one slope is near zero ($|\text{slope}| < 0.1\,\mathrm{m\,yr^{-1}}$), so the apparent sign switch mainly reflects near-stable conditions in one window rather than a clear reversal. Only ~ 5% show clearly inconsistent opposite-sign trends with $|\text{slope}| \geq 0.1\,\mathrm{m\,yr^{-1}}$

in both time windows. Agreement is similar at the scale of IPCC regions ($\rho_s = 0.58$; 78% consistent), although a small number of regions exhibit sign changes in median trends (notably EAU, as well as NCA, SAS, and MDG in Figure 5). Large trend magnitudes (i.e., strongly upward or strongly downward trends, $|\text{slope}| > 0.5\,\text{m}\,\text{yr}^{-1}$) occur less frequently over 19-year windows than over 9-year windows. This is consistent with longer records integrating low-frequency variability and episodic events (Baulon et al. 2022b; Kuss and Gurdak 2014).

For subsequent analyses, we use a combined trend product, assigning 19-year trends where available and otherwise using 9-year trends. Globally, 24% of observations show significant GWL trends, of which 54% are downward. GWL changes are mapped with their magnitude and direction in Figure 5. The prevalence of declining and rising GWLs varies spatially, ranging from under 10% (e.g., northern Europe – NEU) to over 40% (e.g., South Africa – WSAF and ESAF; North and Central America – WNA, CAN, and NCA). Individual observations reveal strong within-region variability, with localized clusters of opposing trend directions and magnitudes.

3.5 Trends in rural and urban coastal regions

Urban and rural areas often differ in land use and water management, which can influence groundwater dynamics. We therefore stratify the dataset by a rural–urban typology (Eurostat, 2021). Figure 5 reveals rural–urban differences in GWL changes. Overall, rural areas show a slightly higher prevalence of significant GWL trends ($|\text{slope}| \geq 0.1\,\text{m}\,\text{yr}^{-1}$) than urban areas, and the trend magnitude is weakly negatively correlated with population density ($\rho_s = -0.17$). This global rural–urban contrast is driven mainly by specific regions, including India (SAS), northern Central America (NCA), Australia (SAU, EAU, and NAU), and New Zealand (NZ). Urban observations show more frequent trends in fewer regions, especially southern Central America (SCA), central North America around the Gulf of Mexico (CAN), and eastern South Africa (WSAF). Overall, these regional differences sum to a higher global prevalence of significant trends in rural than urban areas (26% vs 22%), with declines accounting for ~ 54% and ~ 55% of observations with trends, respectively.

3.6 Identifying hotspots of seawater intrusion susceptibility

To link hydroclimatic susceptibility from aridity and land–sea hydraulic gradients (clusters C1-C4) with dynamic change, we examine how GWL trends vary across clusters and coastal distances. First, we quantify pairwise relationships using Spearman correlations between absolute trend magnitude (|slope|) and aridity, WTD, and hydraulic gradient. These correlations are stronger inland than near the coast. At 100 km from the coastline, the trend magnitude correlates with drier conditions ($\rho_s = 0.56$), with deeper water tables ($\rho_s = 0.63$), and with larger land–sea hydraulic gradients ($\rho_s = 0.36$). Aridity and WTD co-vary ($\rho_s = 0.50$), indicating that drier conditions tend to coincide with deeper water tables. At 1 km from the coast, associations weaken but remain detectable for aridity and WTD ($\rho_s = 0.19$ and $\rho_s = 0.41$), whereas the association with hydraulic gradient is negligible. Second, consistent with these pairwise patterns, trend frequencies differ across susceptibility clusters (Figure 6). Water-limited clusters (C1 and C2), which often coincide with deeper water tables in arid regions, show the highest frequencies of both upward and downward trends across all coastal distances (absolute trends: 36% and 34%). Although less susceptible in our gradient–aridity screening due to steep gradients, C2 shows the highest prevalence of GWL declines that are most pronounced further inland. Conversely, overall trend activity is much lower in energy-limited clusters (C3 and C4; 3% and 12%). Near the coast, susceptibility and change coincide most strongly: within 1 km, 28% of C1 observations show significant trends, with marginally more GWL declines than rises.

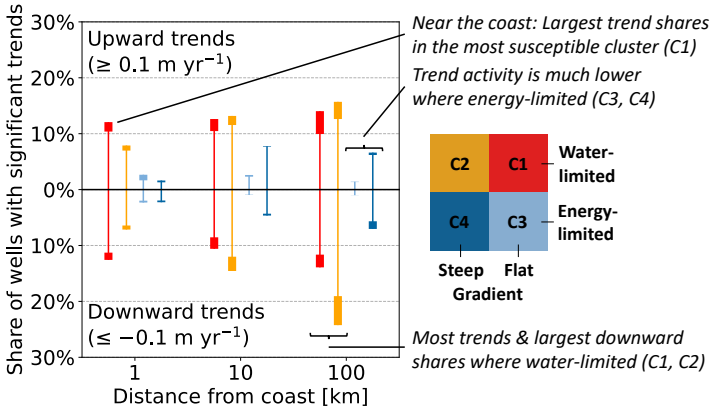


Figure 6: Share of wells with significant upward (top, $\geq 0.1\,\mathrm{m\,yr^{-1}}$) and downward (bottom, $\leq -0.1\,\mathrm{m\,yr^{-1}}$) trends as a function of distance from the coast (1, 10, 100 km). Thicker bar segments at the top and bottom indicate the shares of strongly rising ($> 0.5\,\mathrm{m\,yr^{-1}}$) and strongly declining ($< -0.5\,\mathrm{m\,yr^{-1}}$) observations, respectively. Differently colored shares refer to the SWI susceptibility clusters (C1–C4).

Most trend slopes are moderate ($|\text{slope}| < 0.5\,\text{m}\,\text{yr}^{-1}$), yet even modest head changes can shift land–sea gradients toward or away from flat-gradient conditions. We therefore used derived GWL trends to extrapolate whether observations may enter, remain in, or exit the flat-gradient susceptibility clusters (C1/C3) over the next decade (Figure 7a). Persistent gradient-based hotspots (93.4% of all sites) are locations already exhibiting SWI-susceptible conditions that are projected to remain within C1 or C3 over the next decade. A smaller fraction (3.1%) is projected to stabilize, meaning they may transition out of C1/C3 if recent upward trends continue, and 3.5% are emerging hotspots, i.e., locations projected to transition into flat-gradient conditions (C1/C3) over the next decade.

3.7 Discussion

The results of this study are based on deriving physically motivated indicators of SWI susceptibility from integrated in situ GWL observations. Our hydroclimatic susceptibility framework links land–sea hydraulic gradients with aridity conditions and incorporates recent GWL changes to assess how susceptibility states may evolve, even when detailed hydrogeological information is limited. It methodologically extends elevation-based screening approaches (Jasechko et al. 2020) and complements previous large-scale GWL trend syntheses (e.g., Chávez García Silva et al. 2024; Jasechko et al. 2024) by focusing explicitly on coastal aquifer settings. In the United States, coastal regions that we classify as hydroclimatically susceptible show strong spatial concordance with areas highlighted by Jasechko et al. (2020): within 10 km of the coastline, 20% of our observations have groundwater elevations below sea level (versus 27% of wells in Jasechko et al. 2020), and 37% have flat gradients ($\leq 10^{-3}$). Globally, our flat-gradient, water-limited cluster (C1) captures major regions with documented SWI compiled by Cao et al. (2021) and Kretschmer et al. (2025a) (Figure 3 and Figure 1 in the respective publications). These include the southeastern coasts of the U.S., Central America, the Mediterranean, the Cape Town area, parts of India, and southeastern Australia, which also show a higher prevalence of downward GWL trends in our analysis. Beyond these documented hotspots, our gradient–aridity framework maps the global extent and relative frequency of comparable hydroclimatic configurations, providing a quantitative screening to guide where monitoring and management could be prioritized. This also helps mitigate the urban bias noted by Kretschmer et al. (2025a) because our analysis draws on routine monitoring observations across both urban and rural settings, whereas published SWI case studies are disproportionately from urban, economically developed regions (Kretschmer et al. 2025a). At the same time, the availability of GWL observations in our analysis shows

a heterogenic spatial distribution globally. Coverage is strongest in temperate regions such as Europe and the United States, whereas most of Africa, equatorial regions, large parts of South America and Asia, and many small islands remain sparsely represented or are completely missing. Accordingly, several SWI sites reported in Asia, Africa, and South America (Kretschmer et al. 2025a; Cao et al. 2021) fall outside our analyzed dataset. Our SWI susceptibility patterns therefore reflect the spatial footprint of available observations, and the absence of mapped hotspots in data-poor regions should not be interpreted as the absence of SWI susceptibility.

This study further shows that continued GWL change can generate SWI-relevant gradients within a single decade, and that GWLs near the coastline tend to change most strongly where flat hydraulic gradients coincide with water-limited conditions (C1; most susceptible). While our observation-based indicators do not quantify volumetric freshwater losses, even small deviations and variations around the gradient threshold may become critical over longer timescales. Rising GWLs are as frequent as declines, both globally and often regionally. Coastal aquifers are therefore not uniformly moving toward greater SWI susceptibility. Instead, susceptibility can strengthen or weaken locally, and short-term variability in GWLs, even without sustained declines, can temporarily shift systems toward flat-gradient, SWI-prone conditions. This spatial and temporal variability highlights the value of sustained GWL observations for locally grounded detection of SWI-susceptible conditions.

Further increases in the marine boundary head are likely to shift additional coastlines into SWI-susceptible gradient states (Zamrsky et al. 2024; Adams et al. 2024): The global mean sea-level is currently rising at ~ 4.5 mm yr^{-1} (Hamlington et al. 2024), and is projected to rise by roughly 0.3–1.0 m by 2100 (relative to 1995–2014, depending on emissions scenario), with possible multi-meter rise on longer timescales as peak warming increases (Kopp et al. 2023). This situation can be further exacerbated by pumping-induced upconing, which can cause SWI even further inland (Jakovovic et al. 2011). This matters because our study shows that deep aquifers in rural, arid regions exhibit more widespread declines. Because these systems often respond slowly, management interventions may take years to decades to translate into measurable head recovery, so expectations for SWI susceptibility and monitoring strategies need to account for delayed results. Furthermore, the spatial distribution of significant GWL changes does not align with population density. This should be read as an observed GWL signal, not a direct proxy for local groundwater use, because urban demand is partly met by pumping outside cities. However, our results align with Jasechko et al. (2024), who found widespread declines in dryland aquifers, particularly under croplands. Hence,

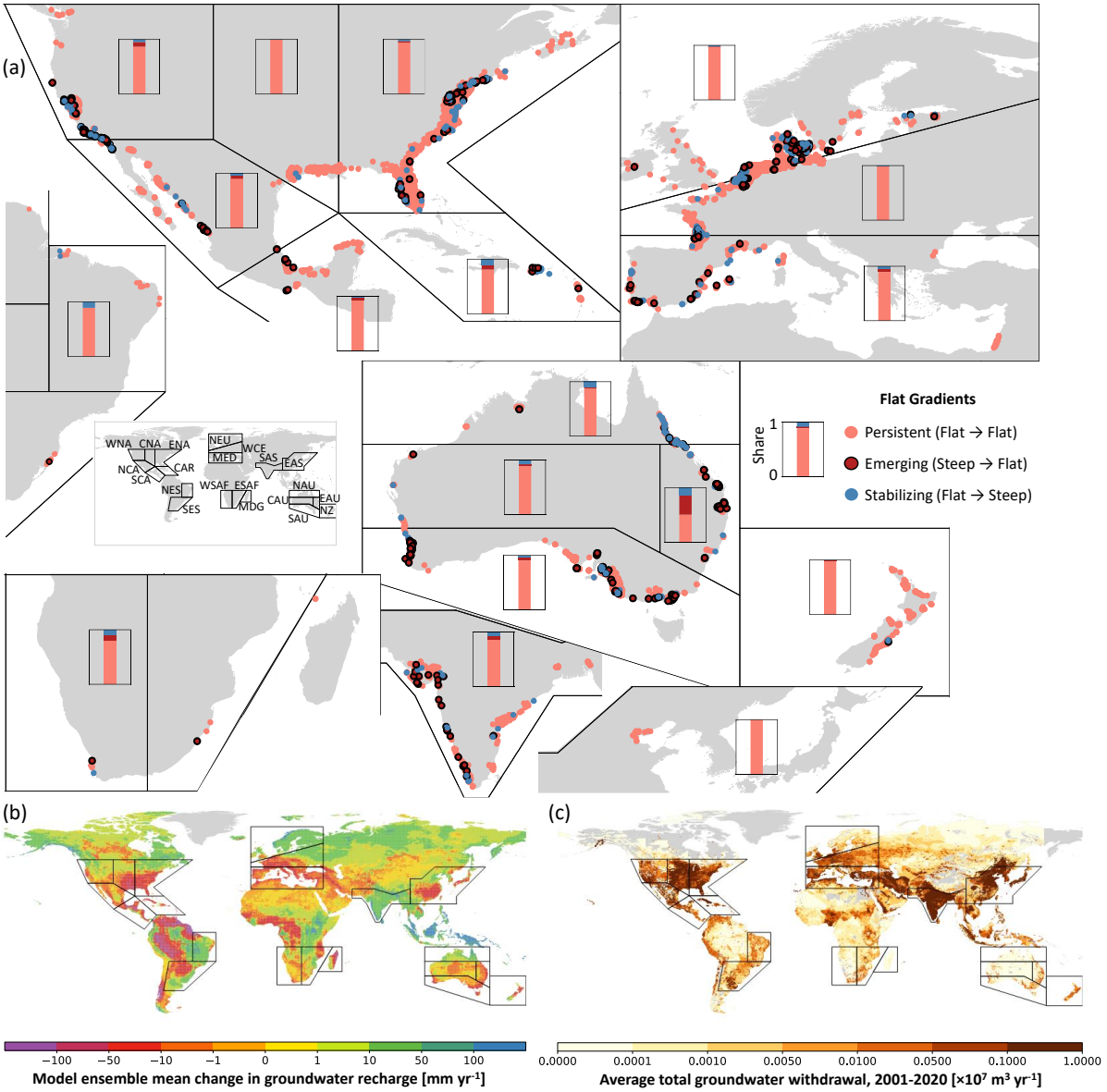


Figure 7: (a) Spatial patterns of hydroclimatic, gradient-based SWI-susceptibility hotspots across IPCC regions (Iturbide et al. 2020), derived from trend extrapolation. Points show individual observations (grid cell centers). Hotspots are defined relative to the flat-gradient clusters C1/C3. When co-located, emerging (entering C1/C3 within 10 years) and stabilizing (exiting C1/C3 within 10 years) hotspots are randomly overlaid on persistent hotspots (in C1/C3 now and in 10 years). Bar charts show the relative proportions of hotspot types within each region ($\geq$ 10 observations). (b) Model-ensemble mean change in groundwater recharge (mm yr^{-1}) between present-day warming ($\sim$ 1 °C) and 3 °C global warming, from the ISIMIP2b multi-model ensemble (Reinecke et al. 2021). (c) Long-term mean total groundwater withdrawal (2001–2020; $\times 10^7\,\mathrm{m}^3\,\mathrm{yr}^{-1}$), from the data-driven Global Groundwater Withdrawal (GGW) model (Nazari et al. 2025).

groundwater-dependent agricultural regions may face disproportionate SWI susceptibility where high extraction pressures coincide with limited regulatory oversight. Persistent or emerging flat-gradient patterns in our study frequently co-occur with regions where sectoral groundwater withdrawals are already high (Figure 7c), with agriculture as the dominant groundwater user and increasing withdrawals in many regions since 2001 (Nazari et al. 2025). Beyond SWI, some regions with relatively low gradient-based susceptibility, such as parts of South Africa, show substantial GWL declines together with projected groundwater recharge reductions (Figure 7b), indicating emerging groundwater-stress risks that are highly relevant in the context of "Day Zero" drought-driven water-scarcity crises, where prolonged hydroclimatic deficits and high demand can push water systems into acute supply shortfalls (Ravinandrasana and Le Franzke, 2025). In such situations, groundwater can be considered a fallback supply (Nolte et al. 2021), but in coastal settings where our indicators suggest SWI susceptibility and/or declining GWLs, this option may be constrained by limited availability.

3.8 Conclusions

This study provides the first global observation-based assessment of hydroclimatic conditions associated with SWI susceptibility in coastal groundwater systems that is based on in situ GWL observations from ~ 480,000 coastal monitoring locations. By combining land–sea hydraulic gradients, aridity, and recent GWL trends in a bottom-up approach that captures spatiotemporal signals often missed by static frameworks and/or coarse top-down approaches, the study delivers three main contributions. First, it translates the available monitoring footprint into globally comparable quantities that can inform large-scale assessments and reporting (e.g., by international frameworks). Second, the study maps where potentially susceptible conditions occur and where they are changing, helping prioritize monitoring and management in coastal regions where risk may be shifting in real time. Third, it provides observation-based indicators that can be systematically incorporated into SWI risk assessments and used as constraints to evaluate and improve large-scale models that simulate unmonitored coasts. At the same time, the hydroclimatic susceptibility indicators do not by themselves resolve local vulnerability or control-response relationships. Doing so requires more detailed hydrogeologic characterization, co-located salinity monitoring, and improved global datasets of key controls, alongside broader efforts to expand and strengthen groundwater monitoring, particularly in regions with large observational gaps. Avoiding or reducing SWI in susceptible settings is critical because salinization of groundwater threatens safe drinking-water supplies, lowers agricultural productivity, and degrades coastal and nearshore ecosystems. Looking

ahead, our findings highlight that stronger monitoring and data-synthesis efforts can enable scalable, observation-based tools to detect emerging SWI risk early, improve system understanding, and inform timely adaptation. Such advances are increasingly important as projected shifts in groundwater recharge, continued sea-level rise, and anthropogenic pressures interact with existing flat-gradient settings in many coastal regions.

3.9 Methods

We used the Coastal GroundWater Level (CGWL) dataset, a newly compiled global resource of in situ coastal GWL observations from governmental monitoring networks and public portals (Chapter 2). The CGWL dataset comprises ~ 629,000 well records (~ 480,000 unique locations at 3-decimal coordinates) from wells within 100 km of the coastline and with WTDs shallower than 100 m (Appendix A.1). Raw groundwater data were quality controlled and harmonized to consistent groundwater depth (WTD; ground-surface reference) and groundwater elevation, using source-provided elevations where available and otherwise CoastalDEM (Kulp and Strauss, 2019) or SRTM (NASA JPL, 2020). We aggregated groundwater information to hexagonal grid cells with a mean cell area of 0.74 km² (H3-8; Uber Technologies Inc 2024) and refer to these grid-cell aggregates as "observations".

3.9.1 Groundwater hydraulic gradients

Land–sea hydraulic gradients were calculated at the grid level using median groundwater elevation and the median Euclidean distance of wells in the cell to the coastline (dataset: Sayre et al. 2021). Using medians reduces sensitivity to outliers but also smooths short-term extremes in the hydraulic head that can be relevant for SWI. Because this proxy does not resolve time-varying coastal boundary heads, non-linear flow paths, or within-cell hydrogeologic heterogeneity, gradients may be biased high or low at individual locations. We therefore interpreted them as a first-order, observation-based screening diagnostic of prevailing coastal head conditions. The sample size for gradient analysis comprised 232,419 grid cells with available groundwater elevation and coastline distance data.

3.9.2 Trend analysis

To quantify recent GWL change, we filtered monthly CGWL time series and computed monotonic trends over 9- and 19-year windows within 1990–2024. The record lengths

used for the trend analysis were chosen iteratively to balance temporal robustness against global support, consistent with prior large-scale groundwater trend studies (Jasechko et al. 2024). A detailed discussion of window-length trade-offs is provided in the Appendix B.3. Time-series selection further allowed variable sampling while limiting seasonal distortion (Appendix B.1). This yielded 51,263 unique monitoring locations with sufficient data to estimate a 9-year trend. Of these, 20,857 locations also met the minimum length for a 19-year trend. Where both are available, the 9-year window corresponds to the final 9 years of the 19-year window. Monotonic trends in GWLs were calculated using the non-parametric modified Mann-Kendall test (Yue and Wang, 2004), and corresponding trend magnitudes were estimated using Sen's slope (Sen, 1968). Trends were computed at the well level and subsequently aggregated to H3-8 grid cells by taking the median of slope values, regardless of individual statistical significance ($p > 0.05$), however, significance flags were retained to support additional filtering. Trend slopes were classified into five trend categories: strongly downward ($< -0.5\,\mathrm{m\,yr^{-1}}$), moderately downward ($-0.5$ to $-0.1\,\mathrm{m\,yr^{-1}}$), no trend (between -0.1 and $0.1\,\mathrm{m\,yr^{-1}}$ and/or non-significant), moderately upward (0.1 to $0.5\,\mathrm{m\,yr^{-1}}$), and strongly upward ($> 0.5\,\mathrm{m\,yr^{-1}}$). Only a small fraction of results classified as "no trend" are non-significant solely due to statistical insignificance (5.6% for 9-year trends and 1.7% for 19-year trends). After aggregation, trend results were available for 36,579 grid cells (9-year) and 17,703 grid cells (19-year), so that 19-year trends support ~ 41% of grid-level trend estimates globally (Appendix B.2).

3.9.3 Hydroclimatic classification and hotspot extrapolation

Hydroclimatic susceptibility was applied along two axes: aridity (as a proxy for recharge limitation) and land–sea hydraulic gradient (as a proxy for the tendency toward seaward discharge versus inland salinity advance). To classify gradients, we used two categories based on model-informed thresholds: flat ($\leq 10^{-3}$, landward or near-zero seaward gradients) and steep seaward gradients ($> 10^{-3}$), following Ferguson and Gleeson (2012) and Kretschmer et al. (2025b), who showed that gradients lower than 10^{-3} are more prone to SWI or to the persistence of saline groundwater. Each H3-8 grid cell was further assigned the aridity index (AI) from Zomer et al. (2022), which represents the ratio of precipitation to potential evapotranspiration. Cells were categorized as either water-limited (AI≤1) or energy-limited (AI>1). These binary classes define the clusters C1–C4 used throughout this study:

- C1 (most susceptible): Flat gradient ($\leq 10^{-3}$) & water-limited (AI≤1)

- C2: Steep gradient ($> 10^{-3}$) & water-limited (AI$\leq$1)
- C3: Flat gradient ($\leq 10^{-3}$) & energy-limited (AI>1)
- C4 (least susceptible): Steep gradient ($> 10^{-3}$) & energy-limited (AI>1)

To analyze potential hotspot persistence/emergence under continued change, we extrapolated groundwater elevation using the estimated trend slope over a 10-year horizon, recomputed the projected gradient, and evaluated whether grid cells would remain in, enter, or exit flat-gradient conditions (C1/C3). This provides a straightforward diagnostic that does not account for potential nonlinearities in drivers of GWL trends or management responses.

3.9.4 Preparing secondary datasets for contextual analyses

The global aridity index (Zomer et al. 2022) is used directly in the SWI susceptibility classification. All other secondary datasets were compiled only to contextualize and interpret gradients and GWL trends. Specifically, Figure 7b-c shows gridded sectoral groundwater withdrawals (Nazari et al. 2025) and a multi-model ensemble of groundwater recharge change (Reinecke et al. 2021) for visual comparison with our results. We prepared the population density of the year 2020 (CIESIN 2018) to derive rural–urban classes following Eurostat (2021). This dataset was preprocessed in ArcGIS Pro (Esri 2022) where information was derived for H3-8 grid cells using zonal statistics.

3.9.5 Limitations

- **Spatial aggregation framework:** We adopted the regular, globally near-equal-area H3-8 grid (Uber Technologies Inc, 2024) as a transparent, assumption-light spatial unit that provides area-representative summaries while reducing the mixing of heterogeneous aquifer conditions relative to coarser aggregation units. This choice reflects that, at the global scale, aquifer boundaries and vertical structure cannot be delineated consistently from the available data in the CGWL dataset and existing global hydrogeologic datasets (Appendix A.2). We used H3-8 grid cells as the primary analysis unit, but repeated all key diagnostics on coastal HydroBASINS level-12 catchments (HYBAS-12; Lehner and Grill 2013) as a sensitivity test. Aggregation error, quantified as the root mean squared error (RMSE) between well-level and unit-level median WTD, is lowest for H3-8 and increases for larger

aggregation units, while it decreases primarily with the number of wells per unit. HYBAS-12 basins are on average ~ 141 km^2 (about four times larger than H3-6 hexagonal grid cells, ~ 36 km^2), yet the share of multi-well units is similar for HYBAS-12 and H3-6, and aggregation error is slightly larger for HYBAS-12, indicating no systematic reduction of aggregation error relative to grid-based schemes with comparable well support (Appendix A.3). We therefore assume that any discrepancies between H3-8 and HYBAS-12 primarily reflect aggregation and sampling effects, that is, how differently sized spatial units mix heterogeneous conditions and how unevenly well density weights regional summaries. Against this background, HYBAS-12 closely reproduces the H3-8 patterns across IPCC coastal regions and distance-to-coast bands for flat-gradient prevalence, trend activity ($|\text{slope}| \geq 0.1\,\text{m}\,\text{yr}^{-1}$), the fraction of downward trends among trend-active units, and the share of persistent/emerging flat-gradient hotspots (Spearman $\rho_s = 0.87$–0.95). Global differences are small, although a few regions deviate by >10 percentage points. For example, grids yield higher flat-gradient prevalence in Western Europe (WCE), while HYBAS-12 shifts the Caribbean (CAR) toward more upward and fewer downward trends, resulting in fewer persistent/emerging hotspots (Appendix B.4).

- **Bottom-up approach and representativeness limits:** Our analysis is global in extent but bottom-up, scaling local well observations to consistent spatial units, and from there to regional and global SWI susceptibility indicators. The spatial representativeness of these indicators, both geographically and with respect to hydrogeologic settings, is constrained by uneven data coverage globally and the limited hydrogeologic characterization (notably aquifer system architecture and confinement). Our chosen fine grid-based aggregation can overweight densely monitored coastal areas and may regionally yield slightly different hotspot and trend-activity shares than coarser units, but it reduces the mixing of heterogeneous aquifer conditions.

- **Simplifications with GWL-based indicators:** While this study uses GWL-based indicators to assess SWI susceptibility, it necessarily simplifies the physical complexity of coastal aquifer systems, where salinization can propagate through multiple interconnected pathways: over land (storm surge and tidal flooding), under land via groundwater transport, and through natural and engineered surface-water networks, and via vertical processes such as upconing of deeper saline water beneath pumping wells (Helton et al. 2025; Werner et al. 2012). Our indicators do not explicitly represent salinity or density-driven flow, limited by salinity observations

being far less available than GWL time series at large scales (Panthi et al. 2022; Thorslund and van Vliet 2020) and co-located water-quality and water-quantity records being uncommon (Ebeling et al. 2022). The indicators also simplify temporal dynamics relevant to SWI. In particular, gradients often rely on a single or short snapshot of head data, which is insufficient to quantify variability or detect shifts, while monotonic trends summarize net change over the analysis window and may miss transient or nonlinear changes.

- **Trend robustness and record-length sensitivity:** Since trend direction at individual locations can be sensitive to record length and to the timing of multiannual variability or management interventions, regions with spatio-temporally inconsistent trends require particularly careful interpretation. However, our 9- vs 19-year comparison indicates that, while longer records are generally more robust, 9-year time series already capture the qualitative direction of GWL change for most regions, with notable exceptions such as in southeastern Australia and India, where declines are stronger and more frequent in the more recent 9-year window. Multi-window trend reporting can serve as a practical robustness check when long GWL records are unavailable.

- **Limits to causal inference and attribution:** In general, this study does not establish causation. Because gridded population products redistribute census counts, population-based contrasts (and any future projections) carry spatially heterogeneous uncertainty. Specifically, they can systematically underrepresent rural populations and misallocate density within cities (Láng-Ritter et al. 2025; Kuffer et al. 2022). Given the substantial uncertainties in both global recharge projections and pumping estimates when interpreted at the scale of local observations (Reinecke et al. 2021; Nazari et al. 2025), we do not directly attribute the observed gradients and trends to these drivers (Figure 7b-c). Instead, we use the aridity index as a proxy for recharge limitation and, indirectly, for potential future increases in water stress and demand. However, we acknowledge that its bi-categorical application (like the gradient split) is a coarse simplification that facilitates transparent global screening, but can mask precipitation effects relevant to recharge and compress diverse hydro-climatic (and hydro-ecological) settings into broad classes (Taylor et al. 2013; Zomer et al. 2022).

4 Disentangling coastal groundwater level dynamics in a global dataset

4.1 Abstract

GWL dynamics result from a complex interplay between groundwater systems and the Earth system. This study aims to identify common hydrogeological patterns and to gain a deeper understanding of the underlying similarities and their link to physiographic, climatic, and anthropogenic controls of groundwater in coastal regions. The most striking aspects of GWL dynamics and their controls were identified through a combination of statistical metrics, calculated from about 8,000 groundwater hydrographs, pattern recognition using clustering algorithms, classification using random forest, and SHapley Additive exPlanations (SHAPs). Hydrogeological similarity was defined by four clusters representing distinct patterns of GWL dynamics. These clusters can be observed globally across different continents and climate zones but simultaneously vary regionally and locally, suggesting a complicated interplay of controlling factors. The main controls differentiating GWL dynamics were identified, but we also provide evidence for the currently limited ability to explain GWL dynamics on large spatial scales, which we attribute mainly to uncertainties in the explanatory data. Finally, this study provides guidance for systematic and holistic groundwater monitoring and modeling and motivates a consideration of the different aspects of GWL dynamics, for example, when predicting climate-induced GWL changes, and the use of explainable machine learning techniques to deal with GWL complexity – especially when information on potential controls is limited or needs to be verified.

4.2 Introduction

When GWL dynamics are tracked over time using groundwater hydrographs, the quantitative status of groundwater resources can be determined. Groundwater resources are exploited and measured locally, and GWL dynamics are subject to processes in their immediate local and regional environment (Bear 1979), such as groundwater recharge rates, groundwater flow and pumping, and SWI in coastal aquifers (Costall et al. 2020; Parisi et al. 2023). Hence, they are strongly location-bound, and there is "a need for groundwater assessments at the field level" (United Nations 2022). To this end, it is common practice in local studies to incorporate direct information from the groundwater system combined with expert knowledge of potential controls in numerical, statistical,

or machine learning models (e.g., in Knowling et al. 2015; Güler et al. 2012; Lee et al. 2019). However, comprehensive knowledge of aquifer processes at the local scale is often lacking, posing even greater challenges for regional, continental, or global assessments of groundwater systems. Such larger scales are important, for example, with regard to large or transboundary aquifer systems, global virtual water trade, and international frameworks such as the UN Sustainable Development Goals (Donnelly et al. 2018; Nimmo et al. 2021). A systemic understanding of aquifer properties is the key to sustainable groundwater management and governance (Guppy et al. 2018; Elshall et al. 2020). Generalized scientific understanding of processes currently relies heavily on aggregating and upscaling local knowledge of groundwater's interactions with the Earth system, however, it is often lacking in direct observation-based evidence.

While it is largely unclear how GWL dynamics compare at large scales, the classification of GWL time series applied on such scales holds the potential to provide valuable insights into hydrogeological similarity (Barthel et al. 2021). For instance, these insights can prove to be valuable in assessing the coherence of large-scale process-based models by focusing on similarities in patterns and spatial trends rather than solely relying on the magnitude of aggregated errors, effectively mitigating the commensurability problem (Gleeson et al. 2021). In contrast, process understanding is limited when only the long-term mean or trend of GWL dynamics is considered (Lischeid et al. 2021; Gleeson et al. 2021; Baulon et al. 2022b).

The present study provides, for the first time, a classification of GWL dynamics at the global scale and is motivated by (a) a large number of GWL data that are available today – although not yet centralized and freely accessible for many regions of the world (United Nations 2022) – and (b) the advancement of data-driven methods in Earth system sciences which are not only capable of finding patterns unseen by humans but also increasingly capable of explaining them (Reichstein et al. 2019; Shen et al. 2018). Previous studies have successfully applied inductive classification approaches (Olden et al. 2012), synthesizing different aspects of GWL dynamics at local to regional scales to investigate physiographic, climatic, and anthropogenic controls of GWL dynamics (Giese et al. 2020; Haaf et al. 2020; Ascott et al. 2020; Wunsch et al. 2021a; Sorensen et al. 2021; Bosserelle et al. 2023), the function of groundwater in ecosystems (Martens et al. 2013), and runoff processes (Rinderer et al. 2019).

The primary focus of this study is to unveil GWL dynamics on global, regional, and local scales by analyzing local data that are distributed globally. The importance and vulnerability of coastal groundwater, which serves as a vital freshwater resource

for ecosystems and large coastal communities with increasing water demands from groundwater (Oude Essink et al. 2010; Famiglietti 2014; Ferguson and Gleeson 2012; Moosdorf and Oehler 2017), prompt this study to focus on disentangling GWL dynamics in coastal regions. We seek to answer the following research questions:

1. How is hydrogeological similarity observed on a global scale, and what are the implications of scaling for similarity patterns?
2. What are the key controlling factors for GWL dynamic patterns on a global scale, and how do they explain the variations observed at smaller scales?
3. What are the current opportunities for and barriers to deriving generalizations of dynamics–control relationships in groundwater using data-driven approaches?

The basis for answering these research questions is a newly compiled large and diverse dataset of about 8,000 GWL time series of coastal aquifers from five continents. Information from the compiled and preprocessed time series (Section 4.3.1) was analyzed holistically using a set of previously defined statistical metrics from Heudorfer et al. (2019) that were computed from individual groundwater hydrographs (Section 4.3.2), thereby reducing their dimensionality (Wang et al. 2006). Section 4.4.1 derives GWL dynamic patterns in coastal regions by using unsupervised clustering techniques (Section 4.3.3) to group GWL time series based on information derived from the aforementioned statistical metrics. To deepen our understanding and elucidate key controlling factors for GWL dynamics, independent data encompassing various potential controls of GWL dynamics and associated surface processes were obtained from global map products (Section 4.3.4) and used to predict clusters in a random forest (RF) classification task (Section 4.3.5) with the results being presented in Section 4.4.2. The RF approach is a robust choice for classifying groundwater dynamics and is widely acknowledged in water science. With its capacity to capture non-linear dependencies and manage uncertainties such as unknown feature importance, overfitting, and outliers, RF provides a reliable tool for exploring complex interactions in natural processes (Tyralis et al. 2019). SHapley Additive exPlanation (SHAP) values were calculated for the RF model (Section 4.3.6). These are used to explain relationships between predicted clusters and descriptive features in Section 4.4.3. Such approaches of explainable machine learning are increasingly and successfully being applied in hydrology (Worland et al. 2019; Yang and Chui 2021; Wunsch et al. 2022; Liu et al. 2022a; Haaf et al. 2023). We discuss hydrogeological similarity results and the scaling effect (Section 4.5.1), evaluate derived explanations for GWL dynamic patterns

(Section 4.5.2), and contextualize our findings within the framework of a case study region (Section 4.5.3). Section 4.6 provides concluding remarks.

4.3 Data and methods

The design of this study consists of the calculation of statistical metrics (hereafter referred to as indices) using compiled GWL time series and the clustering of GWL dynamics by applying unsupervised algorithms to the calculated indices. Subsequently, the clustering result that best differentiates characteristic groups of GWL dynamics was fed into an RF classifier together with local and regional natural and anthropogenic characteristics (hereafter referred to as attributes) from global map products that are potentially related to GWL dynamics. SHAP values were derived to understand the salient controls of each group of GWL dynamics. Unless stated otherwise, data preprocessing, modeling, and visualizations were done using Python (Python Software Foundation 2021).

4.3.1 GWL time series and preprocessing

The GWL time series dataset analyzed in this study was compiled from the national and state governmental agencies listed in Table 1. The data collection spanned from the years 2019 to 2022. Although attempts were made to find data for all coastal regions in the world, data collection focused on regions with a long coastline and large quantities of digitally available data. The wells from the compiled dataset are mainly located in northwestern Europe (Belgium, Denmark, France, Germany, Ireland, Netherlands, and Sweden) but also in Australia, South Africa, Brazil, Canada, and the United States. The wells are part of the GWL monitoring networks of the respective countries, but anthropogenic impact on GWL dynamics cannot be ruled out for all wells. In addition, wells may be affected by SWI and thus have variable density (Post et al. 2007). However, density gradients are not analyzed in this study because our focus is on absolute variations of the GWL. The GWL time series were selected based on the criteria outlined below.

1. There must be an availability of secondary information (in convertible units for homogenization), namely,
   - latitude and longitude of the well location (with variable accuracy),
   - reference vertical point/datum to measure the GWL depth,
   - date of observation.

2. Selection was based on well location and situation.
    - Selection of wells located within 100 km distance from the shoreline. This follows the definitions of Martínez et al. (2007) and Mangor et al. (2017), which focus on ecological, economic, and social importance of coastal water resource planning. Thus, we did not aim to select aquifers directly related to marine processes, such as SWI.
    - Selection of shallow aquifers with mean GWLs less than 100 m below ground surface. It is important to note that this criterion is used as a directional guideline rather than an exact marker as there is uncertainty of up to a few meters in the absolute groundwater depth of observations referenced to sea level that were converted to ground surface reference with elevation data from the Shuttle Radar Topography Mission (SRTM) (Kulp and Strauss (2018) and Rodriguez et al. (2006); source-specific information in Appendix A.6).
3. Selection was based on temporal data availability criteria.
    - Data must be available for at least 4 complete calendar years within the years 1979 to 2019.
    - Time series must not have more than 10% missing observations with data gap lengths of no larger than 2 weeks after aggregation to weekly time steps.

To obtain a homogeneous dataset, further source-specific preprocessing steps were necessary due to the very diverse data presentation, for example, regarding format, labels, and units. The reader is referred to Appendix A.6 for notes on groundwater data collection and preprocessing. Potential anomalies or change points of the individual time series that indicate human activities, measurement, or reporting errors were removed using Density-Based Spatial Clustering of Applications with Noise (DBSCAN; Ester et al. 1996) combined – due to the large number of time series – with visual inspection of suspicious time series because we set the parameters rather conservatively; i.e., with the parameters set, DBSCAN detected outliers in more time series than we would have identified visually (Figure C1). The systematic and additional visual inspection of groundwater time series was found to be beneficial, in line with recent publications (Barthel et al. 2022; Berendrecht et al. 2022; Lehr and Lischeid 2020; Retike et al. 2022). Subsequently, temporal-availability criteria were checked once again, with remaining data gaps being linearly interpolated, and time series were transformed to the 0–1 scale for calculating indices.

4.3.2 Indices

We calculated a total of 45 indices for all GWL time series using an unpublished R package (R Core Team 2021) that was developed with the study conducted by Heudorfer et al. (2019). These indices statistically aggregate and describe various aspects of groundwater hydrographs: their structure (e.g., interannual fluctuation heights and changes), distribution (e.g., alignment of the GWL with upper or lower fluctuation limits), and shape (e.g., steepness of hydrograph increases and decreases). For more detailed descriptions of the used indices (Table C1), we refer to Heudorfer et al. (2019). In addition, there are other promising signature-based time series characterizations for hydroclimatic time series, as selected and explored in Papacharalampous and Tyralis (2022), Papacharalampous et al. (2023), and Wunsch et al. (2021b).

For our approach, the indices were computed from weekly aggregated time series with a length of at least 4 years (Section 4.3.1), which is at the lower end of the time period of 4 to 8 years recommended in Heudorfer et al. (2019). We decided to use short periods to enable a larger pool of time series to be taken into account for the global similarity analysis. The calculated set of indices transforms groundwater information into 45 static metrics. The resulting index value ranges were compared to those of Heudorfer et al. (2019) for quality controlling (Figure C2). As a second outlier treatment (besides applying DBSCAN) of the original time series (see Section 4.3.1), time series with index values in the outermost 0.001% of the index value distribution (4σ) were discarded.

4.3.3 Cluster analysis

Clustering of the indices was performed to find a generalized but robust representation of differing GWL dynamics in coastal regions. No specific indices were selected a priori to capture GWL dynamics holistically. This decision was made because indices differ in their ability to describe GWL dynamics depending on the flow system and the groundwater regime (Heudorfer et al. 2019; Giese et al. 2020; Haaf et al. 2020). For a similar reason, Papacharalampous et al. (2023) recommended using a large variety of time series characteristics for similarity analysis in hydrology.

A principal component analysis (PCA) was applied to index values after standardization to avoid redundancies among the 45 indices (Heudorfer et al. 2019; Figure C3) and to reduce the number of features for clustering. The number of principal components (PCs) to be retained was set based on the scree plot and variance explained, yielding a dimensionality-reduced dataset, where indices are mapped as score values on the significant PCs. This

was followed by a systematic search for the best aggregation of the scores using three different unsupervised clustering algorithms – agglomerative hierarchical clustering using the Ward criterion, k-means clustering, and Gaussian mixture. Unsupervised methods aim to learn more about the internal dependencies among the explanatory variables (Bergen et al. 2019), meaning that no expectations regarding the number or patterns of clusters were pre-set. Instead, evaluation metrics (Rousseeuw 1987; Caliński and Harabasz 1974; Davies and Bouldin 1979) were used to find the best arrangement of data points into clusters via optimizing within-cluster and between-cluster similarity and dissimilarity.

4.3.4 Attribute data

Understanding GWL dynamics requires integrated information from its manifold controls. These are typically represented by attribute data from hydrogeology, climate, land use and land cover, soil and lithology, surface waters, topography, and anthropogenic activity (Moeck et al. 2020; Rajaee et al. 2019; Díaz-Alcaide and Martínez-Santos 2019). For river catchments, such datasets have already been collected for the regional scale (Addor et al. 2017; Klingler et al. 2021) and the global scale, most notably in the HydroATLAS (Linke et al. 2019). However, to date, there are no global datasets available that encompass the above-mentioned attributes specifically for groundwater studies (Haaf et al. 2020). Therefore, attributes mainly describing natural and anthropogenic characteristics and processes at the surface were collected from a variety of recent highest-resolution datasets available at the global scale (Table 2). These were used as proxies for global groundwater-specific datasets in this study.

While river catchments are typically well-defined by topography, subsurface catchments are often complex and vary on small scales (Vahdat-Aboueshagh et al. 2021). Despite recent efforts to develop a similar approach for calculating watersheds with a better representation for groundwater (Nölscher et al. 2022; Huggins et al. 2023), at the moment, there is no best-accepted method for groundwater-relevant delineation of various surface attributes on large scales. One approach to defining the contributing area of a groundwater well is to rely on the immediate vicinity and place a circular buffer around the monitoring site (Johnson and Belitz 2009; Knoll et al. 2019). Using this approach, we extracted attributes as averages from buffers of 500 m radius placed around the well sites. The prepared attributes are of numerical and categorical type and contain information from multiple periods that have been unified where possible. The spatial resolution of the underlying source data is less than 30 arcsec, which is about 1,000 m, except for

hydrological data, which are derived from grids with a spatial resolution of only 1800 arcsec (approximately 55 km). Single- or multidimensional raster data and spatial vector data were downloaded and processed for the well locations as follows (applied to each dataset as indicated in the "Data processing" column in Table 2):

1. Average (median or majority) raster values were extracted for buffer geometries using the "zonal statistics" tool in ArcGIS Pro (Esri 2022) with a cell size of 250 m.
2. Earth Engine Python API was accessed to directly calculate median raster values for buffer geometries and data download.
3. Average (median or majority) values from geospatial vector data were extracted for buffer geometries using the "spatial join" tool in ArcGIS Pro.
4. Distance was calculated between well points and polylines using the "near" tool in ArcGIS Pro.
5. Multidimensional rasters were aggregated to a single raster layer that presents the median or majority of superimposed raster cells of the period from 1981 to the latest available temporal dimension ("make multidimensional raster" tool and "raster aggregation" tool in ArcGIS Pro). Thereafter, we calculated the median of multiple raster files using the "cell statistics" tool in ArcGIS Pro.
6. For the distance of the well location to the closest river, we selected rivers that have a catchment area of at least 10 km^2 or an average river flow of at least 0.1 $m^3 s^{-1}$, or both.

Table 2: List (alphabetically by attribute category) of numeric (num) and categorical (cat) attributes used to relate GWL dynamics to controls and processes in this study.

Category	**Attribute name**	**Short description**	**Unit (converted)**	**Data type**	**Period (averaged)**	**Source temporal resolution**	**Source spatial resolution**	**Reference**	**Data processing***
Anthropogenic	population_density	population density	persons km^{-2}	num	2000	annual	30 arcsec (≈ 1 km)	CIESIN (2018)	1
	irrigation	irrigated area serviced by groundwater	%	num	2000	multi-annual	5 arcmin (≈ 10 km)	AQUASTAT (2021)	1
	groundwater_abstraction	actual net abstraction from groundwater	kg m^{-2} a^{-1}	num	1981–2016	monthly	1800 arcsec (≈ 55 km)	Müller Schmied et al. (2020)	5,1
Climate	temperature_mean	mean annual daily mean air temperature	°C	num	1981–2010	period avg.	30 arcsec (≈ 1 km)	Karger et al. (2017)	1
	temperature_max	highest temperature of any monthly daily mean maximum temperature	°C	num	1981–2010	period avg.	30 arcsec (≈ 1 km)	Karger et al. (2017)	1
	temperature_min	lowest temperature of any monthly daily mean minimum temperature	°C	num	1981–2010	period avg.	30 arcsec (≈ 1 km)	Karger et al. (2017)	1
	precipitation	accumulated precipitation amount	kg m^{-2} a^{-1}	num	1981–2010	period avg.	30 arcsec (≈ 1 km)	Karger et al. (2017)	1
	PET	evapotranspiration of reference crop (ET0)	mm d^{-1}	num	1970–2000	period avg.	30 arcsec (≈ 1 km)	Trabucco and Zomer (2019)	1
	aridity	aridity index (high for more humid conditions, and low for more arid conditions)	–	num	1970–2000	period avg.	30 arcsec (≈ 1 km)	Trabucco and Zomer (2019)	1
Land use and land cover	land use	major land use class	–	cat	1981–2019	annual	30 arcsec (≈ 1 km)	Winkler et al. (2020)	5,1
	ecosystem	major terrestrial ecosystem class	–	cat	–	–	7.5 arcsec (≈ 250 m)	Sayre et al. (2021)	1
Soil and lithology	soc	soil organic carbon 0–5 cm mean	dg kg^{-1}	num	–	–	7.5 arcsec (≈ 250 m)	Poggio et al. (2021)	2,3
	bdod	bulk density 0–5 cm mean	dg cm^{-3}	num	–	–	7.5 arcsec (≈ 250 m)	Poggio et al. (2021)	2,3
	clay	clay content 0–5 cm mean	g kg^{-1}	num	–	–	7.5 arcsec (≈ 250 m)	Poggio et al. (2021)	2,3
	sand	sand content 0–5 cm mean	g kg^{-1}	num	–	–	7.5 arcsec (≈ 250 m)	Poggio et al. (2021)	2,3

Continued on next page

continued.

Category	Attribute name	Short description	Unit (converted)	Data type	Period (averaged)	Source temporal resolution	Source spatial resolution	Reference	Data processing*
	silt	silt content 0–5 cm mean	g kg^{-1}	num	–	–	7.5 arcsec (≈ 250 m)	Poggio et al. (2021)	2,3
	lithology	major lithological class	–	cat	–	–	polygons	Hartmann and Moosdorf (2012)	3
	unconsol_sediment	unconsolidated sediments major class	–	cat	–	–	polygons	Börker et al. (2018)	3
	permeability	surface permeability	log(k) m^2	num	–	–	polygons	Huscroft et al. (2018)	3
Surface waters	streamflow	mean annual river discharge	m^3 s^{-1}	num	1981–2015	annual	30 arcsec (≈ 1 km)	Barbarossa et al. (2018)	5,1
	distance_coast	distance of well location to the coastline	km	num	–	–	polyline	Lehner and Grill (2013)	4
	distance_stream	distance of well location to the closest river	km	num	–	–	polyline	Lehner and Grill (2013)	4,6
	runoff	runoff from land (sum of runoff components)	kg m^{-2} a^{-1}	num	1981–2016	monthly	1800 arcsec (≈ 55 km)	Müller Schmied et al. (2020)	5,1
	TWS	total water storage (sum of compartments)	kg m^{-2}	num	1981–2016	monthly	1800 arcsec (≈ 55 km)	Müller Schmied et al. (2020)	5,1
Topography and sub-surface depths	bedrock_depth	absolute depth to bedrock	cm	num	–	–	7.5 arcsec (≈ 250 m)	Shangguan et al. (2017)	1
	wtd	water table depth	m	num	–	–	30 arcsec (≈ 1 km)	Fan et al. (2013)	1
	elevation	elevation	m	num	1994	–	1 arcsec (≈ 30 m)	Farr et al. (2007)	2,3
	slope	steepness of elevation surface	°	num	1994	–	1 arc sec (≈ 30 m)	Farr et al. (2007)	2,3

*Individual data processing is explained in Section 4.3.4.

4.3.5 Random forest model

The classification task was to assign the correct cluster out of n identified clusters (Section 4.3.3) using a set of attributes (Section 4.3.4). Well locations were considered for the RF classification (Breiman 2001) only when data for all attributes were available. Furthermore, to avoid multiple clusters for a single combination of attributes, the cluster with the highest frequency was selected. The instance was retained if the identified cluster constituted more than 50% of all clusters associated with that specific combination of attributes. The categorical attributes of land use, ecosystem, unconsolidated sediment, and lithology were preprocessed for their use in the RF model using one-hot encoding.

The RF model was optimized and evaluated with a stratified split of the dataset into a group for training and optimization (80% of the data) and a group for testing (20% of the data), taking into account an imbalanced dataset, where not all clusters have equal numbers of wells assigned and where regions are not represented equally by the dataset. For model optimization, hyperparameters were tuned within a 5-fold cross-validation framework. RF is a tree-based method that contains structures for feature selection (Breiman 2001). No explicit feature selection was applied during preprocessing or model optimization, driven by the objective of exploring the relative importance and effects of all potential controls, as well as their collective impacts on GWL dynamics. Evaluation during model optimization and with the testing data of the optimized model was carried out using the model accuracy, calculated as the total number of correctly assigned classes divided by the sample number. In addition, the optimized model, i.e. the success of the classifier, was evaluated with metrics for the individual clusters: precision, recall, and F1 score. Here, precision is the number of correctly identified wells per cluster divided by all the times the model predicts that cluster. Recall is the number of correctly identified wells per cluster divided by the total number of wells in that cluster. The F1 score is 2 times the multiplication of precision and recall ($2 \times \text{precision} \times \text{recall}$) divided by the sum of precision and recall. RF, in general, works reasonably well when using the default values of hyperparameters in common algorithms (Probst et al. 2019). However, performance improvements can still be achieved by tuning, in particular, the number of trees that should be set high enough for robust results and the number of features considered for each time the tree is split. Up to 2,000 trees were tested in combination with either all features or their square root being considered when looking for the best split using two different splitting criteria (Gini impurity and Entropy: Breiman 2001).

4.3.6 SHAP analysis

SHapley Additive exPlanation (SHAP) values (Lundberg et al. 2020) were used to investigate the feature attributions (feature importance and feature effects) to explain the RF classification. The three-dimensional space of SHAP values for the multiclass problem of this study is given by the number of clusters, the number of samples, and the number of features; i.e., attributes. Larger SHAP values for a specific cluster correspond to a higher probability of the cluster label. SHAP values of the samples that are part of the test dataset were analyzed collectively to explain how attributes impact the classification of the model. This corresponds to a global explanation of the model. For some analyses, SHAP values were combined for the attribute categories listed in Table 2. The SHAP values of individual instances, i.e. well sites, were combined with regional information to discuss the extent to which the generalized GWL dynamics derived through the index approach of this study can be explained in a case study region. When explaining controls of GWL dynamics from SHAP values, it is important to consider the interrelationships between the features. From the point of redundant features, the model is mainly trained on the feature with the most significant information in combination with other selected features. Hence, variables that combine information included in other features are generally ranked higher and reduce the ranking of the other redundant features.

4.4 Results

4.4.1 Identification of groundwater level dynamic patterns

Application of the time series selection criteria from Section 4.3.1 resulted in a set of 8,574 GWL time series for which 45 indices were calculated. After removing time series with extreme index values (Section 4.3.2), a dataset with indices for 7,888 time series remained. Overall, the value ranges for the different indices are comparable to those found by Heudorfer et al. (2019) (Figure C2). Several indices showed a strong linear correlation (Figure C3). From the PCA, five PCs represented 70% of the variability of the linearly combined indices, while the first three PCs already described more than half of the variability (Figure C4). The top five PCs served as input for the three clustering algorithms. These algorithms yielded similar results, where the highest cluster separation was achieved with three or four clusters (CL). The k-means clustering yielded the best cluster separation compared to hierarchical clustering and Gaussian mixture for all three evaluation criteria for cluster separation (Figure C5). Comparing the cluster membership of wells, cluster composition based on k-means clustering is found to be quite similar

to hierarchical clustering for most well sites (Figure C6). The various characteristics of GWL dynamics can be assumed to exhibit complex and non-linear interlinkages in the multidimensional space.

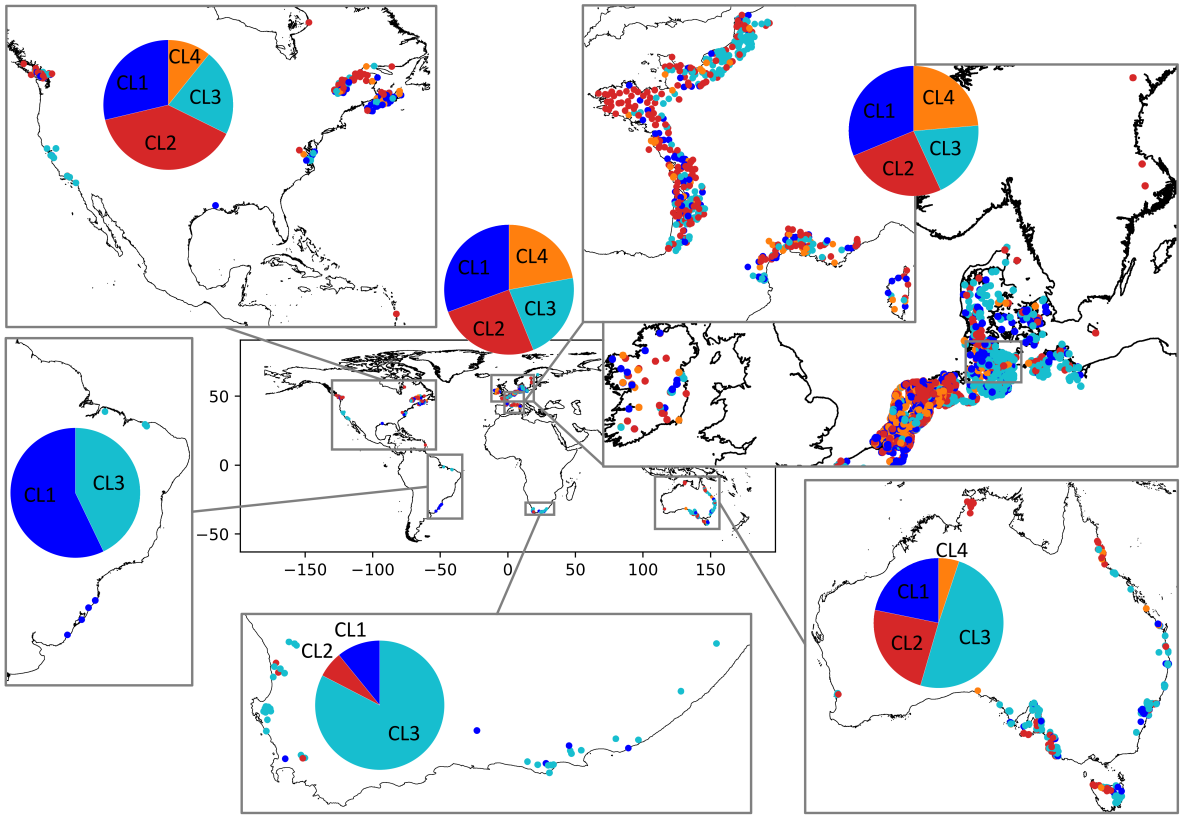


Figure 8: Assignment of groundwater dynamic regimes (clusters) to well locations globally (center) and in coastal regions in North America, Europe (enlarged map of France and northwestern Europe), Australia, South Africa, and Brazil. Due to the large scale, overlapping dots are not jittered, but well points are shuffled to prevent the same clusters from being constantly drawn over others.

Given the nature of the dataset, the success of the different clustering approaches for disentangling GWL dynamics can be mainly attributed to the following two aspects. First, it is plausible that the data distribution may not strictly follow Gaussian distributions because already individual indices show non-linear behavior (Figure C2; Haaf et al. 2020), limiting the effectiveness of Gaussian mixture models. In such cases, the characteristics of the dataset can favor clustering algorithms, such as k-means and hierarchical clustering that, are more flexible with regard to the range of cluster shapes and sizes. Second, multidimensional datasets with complex and non-linear interlinkages may often lack clearly defined and compact clusters (Campello et al. 2020). This becomes evident when analyzing the scatter plot of the first three PCs (Figure C6), making clustering with

k-means and hierarchical clustering challenging. However, despite this limitation, these algorithms still exhibit advantages in terms of capturing the distribution requirements of the GWL dynamics dataset, and k-means clustering disentangles best the main patterns within GWL dynamics. Therefore, k-means clustering was chosen to cluster the indices.

The predominant cluster differs between the coastal regions represented in this study. In regions with a high density of wells, usually all clusters are present (Figure 8). Therefore, similar patterns of GWL dynamics are found in coastal aquifers across the globe. Regional spatial patterns, i.e. geographically concentrated clusters, are visible for South Africa and Australia, where most wells belong to cluster CL3, and within Europe, where most of the wells in the dataset are located. For example, in Germany, wells assigned to CL3 are mostly located around the Baltic Sea and further inland, while the other clusters are pronounced along the exposed west coast. In the Netherlands and Belgium, most wells belong to CL1 and CL2. Most wells in northeastern France are assigned to CL3, whereas most wells in northwestern France are assigned to CL2. Each cluster represents a distinct pattern of GWL dynamics, regardless of shifted or opposing seasons in different regions, as shown in Figure 9. For example, the double hump in the average annual hydrograph of North America, with time series from wells being located mainly in the upper latitudes, could be attributed to snowmelt and is found in all clusters. Patterns are the most precise (smallest confidence intervals) for the European dataset.

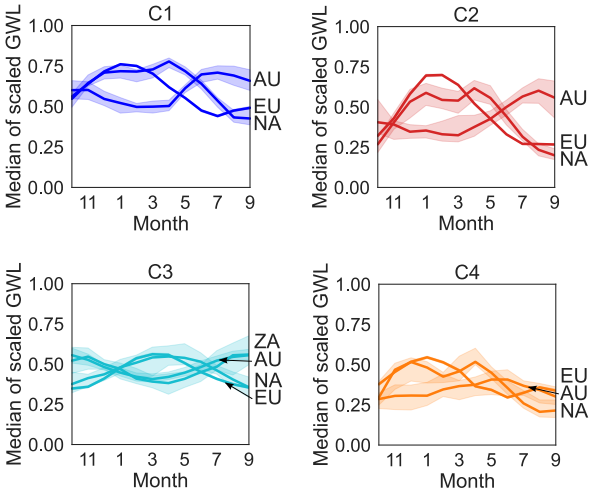


Figure 9: Typical annual hydrographs (median and 95% confidence interval) from averaged groundwater depths of scaled GWL time series for each month per cluster, calculated and plotted for the coastal regions in North America (NA), Europe (EU), South Africa (SA), and Australia (AU) when at least 10 time series associated with the respective cluster were available.

GWLs in three out of four clusters can be located either mostly at their upper (CL1)

or lower (CL4) boundary during the year, with high GWLs in winter and spring to low GWLs in summer, or the GWL fluctuates around its annual median during the year, with high interannual amplitude (CL2). In contrast, GWLs of CL3 are characterized by maximum water levels being reached later in the year.

The clusters can be distinguished with respect to multiple features of the hydrograph, expressed by indices. Thus, the salient pattern of each cluster is defined by several indices and not by an individual one. However, some individual indices allow us to separate clusters or groups of clusters better than the others. Figure 10 shows how different aspects of GWL dynamics, each described by multiple indices (Table C1), contribute to disentangling patterns of GWL dynamics. Many indices of multiple components clearly separate clusters concerning the hydrograph structure, including the regularity of the seasonal amplitude (parde.seasonality) that is small for CL1 and CL3 and largest for CL2, the size of the global amplitude of the unscaled GWL that is comparatively small for CL4, the manifestation of interannual variation (e.g., from base flow stability (BFS) that implies that interannual variation is rather stable in CL3 and unstable in CL4), and different orders of flashiness (e.g., from base flow index (BFI), where the base flow component is large for CL3 and small for CL4). From the time series distribution aspect, GWL dynamics from CL1 can be characterized as upper bounded, while GWL dynamics from CL4 are slightly lower bounded (e.g., from median). Dispersion over the range of GWLs (e.g., from the range of duration curve (fdc.range) between different percentiles) is smallest in CL1 and largest in CL2. The shape of GWL time series in terms of both scale (hydrograph magnitude (sqr.avg)) and slope (e.g., from rise.avg) distinguish the clusters quite well from each other. For example, CL3 is particularly noticeable for combining GWL time series with a small slope, while CL4 combines such with a large slope.

4.4.2 Classification of groundwater level dynamic patterns

In the classification task using RF, controls of GWL dynamics were investigated using more than 5,000 unique cluster–attribute relationships. Stable classification accuracy during hyperparameter tuning (see Section 4.3.5) was achieved with more than 100 trees. Both splitting criteria that were tested performed equally well in terms of classification accuracy, whereas the consideration of different features for the splits led to significantly higher accuracy (Figure C10). The final model of this study was set up using the Gini criterion, randomly varied subsets (square root) of features for each split and 700 trees.

With the final model, approximately 62% of the clusters, on average, were classified correctly; i.e., they were assigned to the correct cluster out of four. Comparing the model's

result accuracy to a scenario where descriptors were shuffled shows an improvement in accuracy of about 60%. By shuffling the attributes, the relationships between the descriptors and the target variable (clusters) are disrupted, and any observed accuracy in the shuffled model is purely due to chance. These results suggest the presence of moderately strong linkages between the attributes and clusters describing distinct GWL dynamic patterns within the RF model. Figure 11 shows that the classification performance is distributed partly unequally across the different clusters. From the confusion matrix, it can be seen that CL3 is most often (78%) predicted correctly and, if not, is either confused with CL1 or CL2, but almost never with CL4. The other clusters are predicted correctly similarly often, with accuracies ranging from 55% to 60%. Recall and precision of the clusters are quite similar. Considering the slightly different support of each cluster, the classification performs on average similarly well in terms of false-positive cluster assignments (e.g., a well site is classified as CL1 when it actually belongs to another cluster), and false-negative assignments (e.g., a well site that belongs to CL1 is not classified as CL1 but as another cluster). Overall, there are only minor differences in classification performance among the various coastal regions.

4.4.3 Model explanations from SHAP values

SHAP values that were calculated for the classification model are used in the discussion to explain the importance and effects of different controls on GWL dynamics, while general results for the RF model are presented here.

By analyzing the correlation of SHAP values, we can detect where feature importance may be affected by information redundancy (Figure C11). In this study, the redundancy among features tends to be small or moderate, and patterns are not very consistent across the clusters. Specifically, the redundancy between feature groups (in correspondence with categories of attributes in Table 2) is minimal for cluster CL3 in the model, whereas it is more evident for cluster CL4. CL2 exhibits redundant contributions from anthropogenic features and features derived from surface waters, with climate features being additionally redundant in the prediction of cluster CL1.

Three groups of features are ranked higher in terms of overall feature importance in the global model compared to the remaining three groups (Figure C11). These are topography and subsurface depths, followed by soil and lithology and climate, with only minor differences in the importance of the features between the clusters. The largest difference is that the feature group of topography and subsurface depths are about twice as important for predicting CL3 compared to the other three clusters. For these clusters,

the feature groups of soil and lithology and climate are similar in importance compared to topography and subsurface depths. Therefore, features of the group of topography and subsurface depths primarily play a role in distinguishing CL3 from the other clusters.

More in-depth insights into how the model can distinguish the clusters – and, in particular CL1, CL2, and CL4 – from each other through the various features can be derived when using SHAP values to link the feature impacts to their effects (Table 3; Figure C12). From the individual features, elevation, slope, and water table depth (wtd) are most important within the feature group of topography and subsurface depths; soil fractions and the (non)occurrence of specific unconsolidated sediments are most important within the feature group of soil and lithology; temperature is most important within the feature group of climate; irrigation is most important within the feature group of anthropogenic; distance to the coast is most important within the feature group of surface waters; and the (non)occurrence of a specific terrestrial ecosystem type is most important within the feature group of land use and land cover.

Category	**Attribute or feature (name)**	**CL1**	**CL2**	**CL3**	**CL4**
Topography and subsurface depths	elevation	<	<	>	<
	wtd	<	<	>	<
	slope	–	<	>	<
Soil and lithology	grain size (clay, silt, sand)	<	>	>	<
	aeolian deposits (unconsol_sediment_Ae)	–	<	<	>
	glaciofluviale deposits (unconsol_sediment_Gf)	–	<	>	–
Climate	temperature (_min; _mean)	–	>	<	–
	precipitation	>	–	<	–
	aridity*	>	>	<	>
Anthropogenic	irrigation	>	–	<	>
	population_density	–	<	<	>
Surface waters	distance_coast	<	–	–	>
Land use and land cover	cold temperate moist grassland on plains	>	–	<	>

Table 3: Qualitative description of the more important and, in their effects, more differentiable features, starting with the feature group with the greatest model importance (Figure C12) and, for the individual clusters, the most differentiating effects (based on Figure C13); < and > mark smaller and larger feature values for one or more clusters compared to one or more other clusters, while there is no clear information for –. *Aridity index: see Table 2.*

The most significant effects that features have on predicting clusters in the RF model are those that best separate clusters from each other (Table 3). CL3 is best distinguished from the other clusters by a comparatively large elevation, large wtd, and large slope. In addition, some effects specifically separate (a) CL3 from CL4, (b) CL4 from CL1 and CL2,

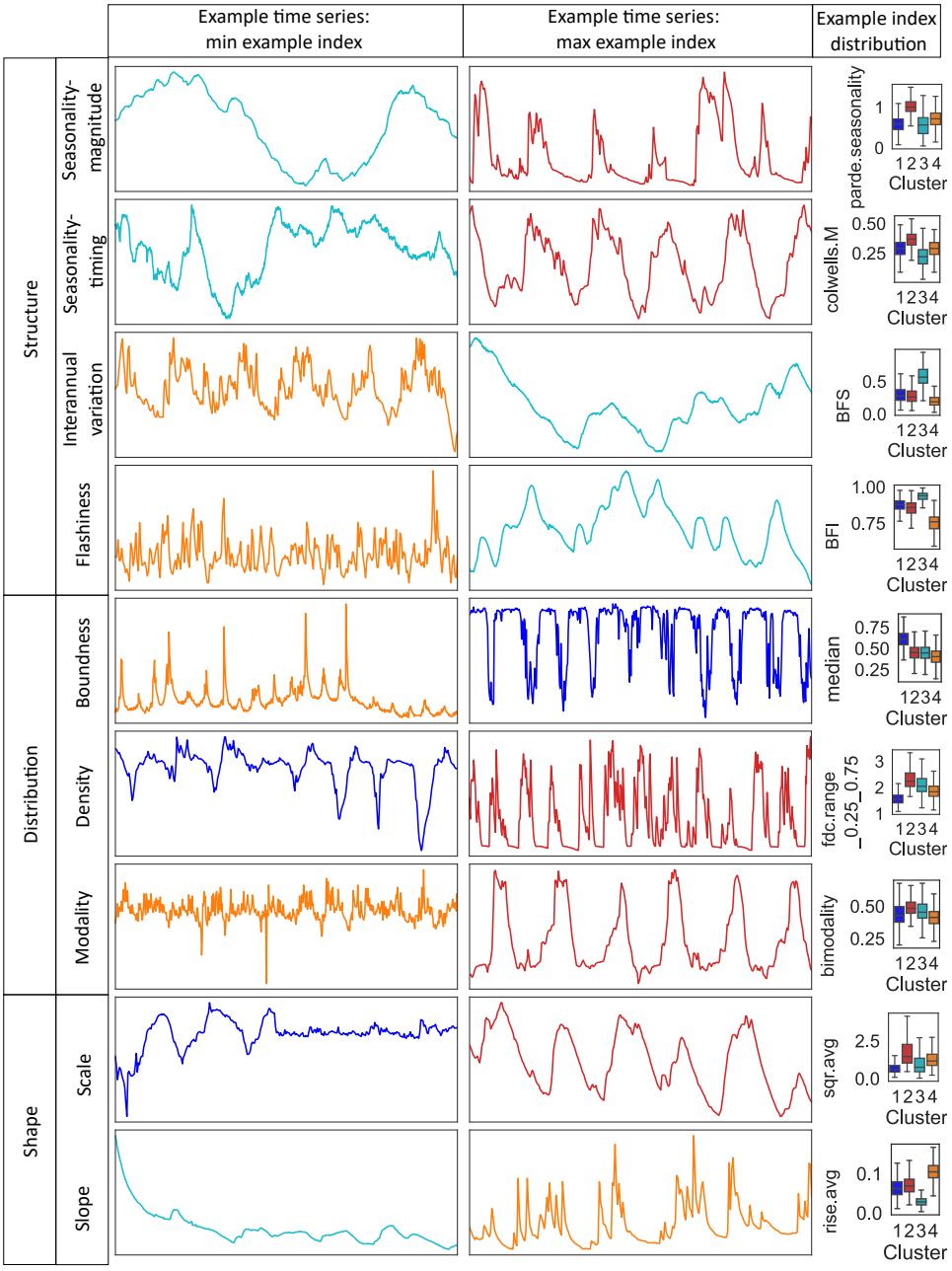


Figure 10: Example groundwater time series from the European continent illustrate typical GWL time series where one aspect from each component of the Heudorfer et al. (2019) typology of GWL dynamics (Table C1) is weak (left column) and one is strong (right column), with an example index from that component that is displayed showing within-cluster variability (see Figure C7 for within-cluster variability for all indices).

and (c) CL1 from CL2, although the latter is the least successful in the model (largest confusion values of CL1 and CL2 in Figure 11). In comparison with CL2, CL3 combines

well locations with lower minimum and average temperatures and sediments typical for glaciofluvial deposits. In comparison with CL1, CL3 combines well locations with less precipitation, more coarse-grained soils, and less anthropogenic activity. Feature effects between CL3 and CL4 are most often contrary, although they are less pronounced for climate. Where CL2 is predicted, there are also comparatively more coarsegrained soils and rarely sediments typical for eolian deposits, and anthropogenic activity is low. These characteristics from soil and lithology, as well as anthropogenic activity, also separate CL4 from CL2. Less differentiates CL4 from CL1. It is noticeable that the distance of the well to the coast is generally higher in CL4 compared to CL1. The typical well locations in CL1 and CL2 differ in their soil fractions (finergrained soils in CL1) and anthropogenic activity (larger in CL1).

Finally, SHAP values for individual wells that are all located in the same region in northern Germany (Figure 12) can be used to understand how the results of the feature importance and effects derived from the global dataset apply regionally. We analyzed the SHAP values of the four well sites from the case study region that have the highest probability of being predicted by the model for each cluster in the test dataset (Figures C14– C17). The wells that are most likely to be predicted as CL1 and CL3 are also included in these clusters, whereas the wells that are most likely to be predicted as CL2 and CL4

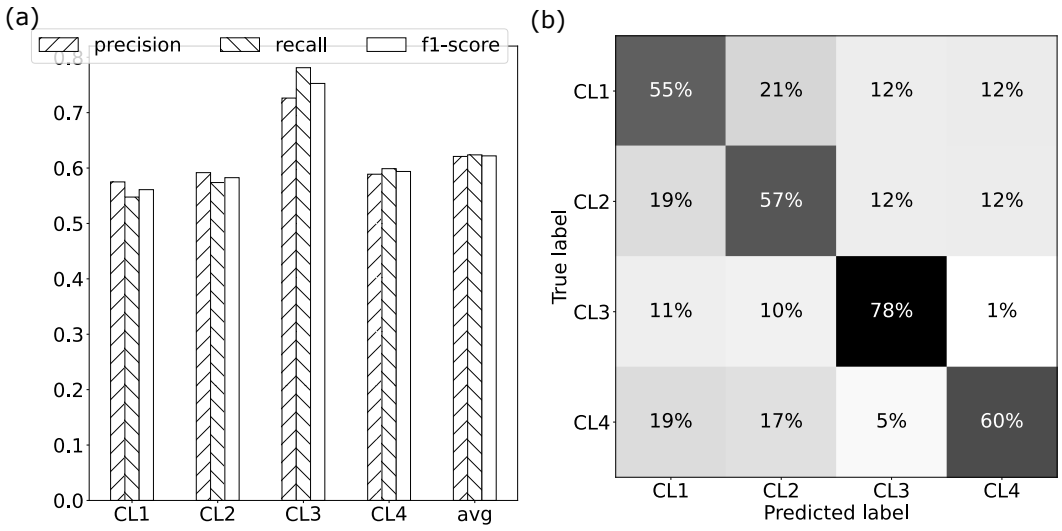


Figure 11: Performance of the classification task using random forest (Breiman 2001). (a) Evaluation metrics of classification (precision, recall, and F1 score) given per cluster and for the weighted average (performance metrics weighted according to the support of each cluster). (b) Normalized confusion matrix. Rows represent the actual cluster, while columns represent the predicted cluster. Thus, the percentages of correctly predicted classes are represented by the diagonal elements, and the confusion (i.e., misclassification) is represented by the off-diagonal elements.

are part of CL3. Thus, although CL3 is overall the best-identified cluster by the model (Figure 11), in the case study region, where most of the wells belong to CL3, it is confused with CL2 and CL4 for the example wells. The correct prediction of CL1 is due to the low elevation and wtd at the well location, as well as the presence of grassland (Figure C14). Due to these characteristics, the well is also not predicted to be CL3, while the region's mean temperatures around 10 °C also favor CL3 at this well site. The statistically averaged temperatures, which should be very constant for the region, can sometimes favor and sometimes disfavor all clusters. According to local feature importance in this region, temperature is the most important attribute following elevation and wtd for the prediction of GWL dynamics types at specific wells. However, when analyzing SHAP values for a model trained on a global dataset, we are primarily examining the global context for a local feature, not the local context. Besides confirming temperature, CL3 is correctly predicted where its prediction probability is largest because of comparably large elevation in the study region and because there is no irrigation (Figure C16). In contrast, a CL3 well is incorrectly predicted as CL2 with low elevation and a population density that is below average for the region (Figure C15). Another CL3 well is incorrectly predicted as CL4 with low elevation and one of the highest clay contents for the region (Figure C17). The described relationships are in line with or not contradictory compared to the global correlations of Table 3.

4.5 Discussion

4.5.1 Hydrogeological similarity and scaling effect

Four predominant types of GWL dynamics were identified for coastal wells distributed across five continents. Our results show that shallow aquifers exhibit more diverse GWL dynamics than deeper aquifers. This suggests that absolute GWL depth is less determinant in differentiating global GWL dynamics than might be assumed based on the clustering of direct time series from a single particular region, as in Rinderer et al. (2019), where, in particular, mean water level and amplitude of variation determine clusters of GWL dynamics.

Based on the spatial distribution of clusters, wells exhibit distinct regional structures that strongly imply the presence of underlying spatial controls, e.g., CL3-type GWL dynamics in northern Germany tend to coincide with Geest land, while CL1-type GWL dynamics are more often found in the marshlands (Figure 12a). Another example is that the topographically elevated Veluwe region in the center of the Netherlands is the only

region at the North Sea where many wells are assigned to CL3 (Figure 8). However, GWL dynamics of different types are also frequently found in wells within a very short distance from each other or even on top of each other (in the case of multiple groundwater storeys). This means that the diversity of GWL dynamics can be larger locally than between two wells in different climate zones on different continents, likely due to the significant influence of heterogeneous hydrological conditions. Therefore, in addition to the findings by Wunsch et al. (2021b), it is shown that nearby wells do not necessarily have a higher degree of similarity in terms of GWL dynamics compared to more distant wells – even when these wells are located on different continents. Regionally similar, but locally diverse GWL dynamics highlight the complexity of generalizing GWL dynamics on different spatial scales. Yet, the fact that our results include spatial patterns at local, regional, and global scales (considering the globally distributed clusters of GWL dynamic types), speaks for the usefulness of our approach in finding generalizations beyond specific geographic contexts. We therefore conclude that our approach is robust with respect to scale dependency. Previously, studies on GWL dynamics have resulted in systemic understanding at local and regional scales (Wunsch et al. 2021b; Giese et al. 2020; Haaf et al. 2020). While this may be more straightforward for applied groundwater management and monitoring at these scales, this study is the first to provide insights into generalizations of GWL dynamics using a global dataset.

We found that using indices for analyzing GWL dynamics is efficient and offers comparability and interpretability. In these terms, this approach might be superior compared to directly analyzing GWL time series or only focusing on the long-term mean or trends. Heterogeneous time series with different seasons, time series lengths, and periods can be combined in this approach with robust results, as previous studies have shown (Heudorfer et al., 2019; Wunsch et al., 2021), thereby reducing the dimension of the time series and potentially better establishing cause-and-effect relationships. From the clustering results, we can see that indices successfully separate between climatic forcing (as present in seasonality) and physiographic forcing, while in practice, it remains challenging to identify what drives GWL dynamics at various scales and settings (Blumstock et al. 2016; Moeck et al. 2020) and to distinguish between the impacts of climatic and other natural factors versus anthropogenic impacts on GWL dynamics and trends (Wriedt 2017; Lischeid et al. 2021). Although indices are also affected by the length of the GWL time series (Heudorfer et al. 2019), they can help to identify time series with "high-amplitude, low-frequency variability," such as in the Normandy (Paris Basin, France; compare Baulon et al. 2022a). Groundwater well sites with dominant interannual variations from multi-annual to decadal scales identified by Baulon et al. (2022a) in this area match well sites with CL3-type

GWL dynamics found in this study (Figure C9). At these well sites, the likelihood of approaching short-term groundwater drought is higher if the prevailing low-frequency variability is associated with a downward phase, while the annual GWL status and the previous winter's groundwater recharge have less influence on short-term drought (Baulon et al. 2022a). Overall, we can therefore assume that longer residence times are typical characteristics for CL3-type aquifers.

The methodological approach underlying the GWL dynamic analysis offers much potential but also has limitations and uncertainties with respect to the validity of the generalizations generated. Despite the fact that general robustness of indices was found starting at time series with a minimum length of 4 years (Heudorfer et al. 2019), uncertainties remain regarding the representativeness of the generalizations for individual wells with this time series period length but less so for the generalizations themselves. Furthermore, Heudorfer et al. (2019) originally defined some of the indices for time series with daily resolution, and decadal GWL dynamics are only indirectly represented by indices in our study (e.g., seasonality in the annual hydrograph). Therefore, further investigation of the effects of different temporal resolutions and time series lengths, as performed in Papacharalampous et al. (2023) for hydroclimatic time series, would be of interest. With respect to the groundwater systems and processes represented by the GWL dynamic patterns, it should be emphasized that there may be a bias in favor of the typical GWL dynamics in northwestern Europe because most of the wells in the dataset are located there.

Classification of specific groundwater processes requires input data specifically related to and known to be influential in relation to these processes. However, there are major limitations with available GWL data, particularly in terms of time series length and data resolution. GWL time series are rarely sufficiently long to recognize the dimension of climate change, making it difficult to establish an effective monitoring and management strategy. High-resolution time series are, for example, required for analyzing the interaction of groundwater with the sea (Haehnel et al. 2024). Accounting for SWI in GWL dynamic pattern analysis is best supported by pattern recognition or correction together with groundwater chemistry data (Narvaez-Montoya et al. 2023; Parisi et al. 2023). However, the expansion of studies on this topic is currently hampered by the lack of GWL time series with high temporal resolution and sufficient lengths, as well as the lack of availability of qualitative groundwater data together with GWL data, among other challenges in groundwater monitoring practice (Rau et al. 2020). Several studies have shown the degradation of groundwater in qualitative and quantitative terms due to its overexploitation (Alfarrah and Walraevens 2018; Peters et al. 2022; Xanke and Liesch

2022), but it was beyond the scope of this study to focus on changes in groundwater systems. Furthermore, more in-depth analyses of the individual indices could help to find out to what extent and in which indices specific processes such as SWI or anthropogenic influence are manifested and thus contribute to a classification.

The index-clustering analysis conducted in this study was instrumental in capturing the major patterns of GWL dynamics of coastal regions worldwide. The utilization of the k-means algorithm has enabled us to represent these patterns at a high level, providing valuable generalizations. However, if in other research questions the necessary level of pattern recognition is smaller, clustering algorithms that do not require both clearly distributed data and clearly defined clusters have the potential to identify additional clusters and uncover more nuanced dynamics within the dataset.

4.5.2 Cause-and-effect relationships

We further analyzed whether the identified types of GWL dynamics are meaningful, i.e., consistent with expected patterns (Yang and Chui 2021), although machine learning can also reveal undetected and previously unexplained patterns. We generally expect patterns as a result of the multitude of natural and anthropogenic factors influencing GWL dynamics. Meaningful GWL dynamics have the potential to derive cause-and-effect relationships between GWL dynamics and their driving forces, a topic for which observation-based evidence is comparatively scarce. Thus, in this study, driving forces were linked to GWL dynamic patterns in an RF classification task. Here, many different environmental attributes describing surface and – to a smaller extent – subsurface processes potentially associated with groundwater recharge and discharge were used to explain (dis)similarities within and between the identified patterns.

While comparing individual aspects of the hydrograph with potential controlling factors can only explain specific aspects of GWL dynamics (e.g., Haaf et al. (2020) related flashiness (BFI) to focused recharge as a consequence of depressions or connectivity to streams), the holistic analysis of various GWL dynamics aspects against controls enables the estimation of the influence of multiple controls over processes defining groundwater quantity over time.

Since GWL dynamics in coastal regions can be similar in different regions globally, similar GWL dynamics are not necessarily the result of the same cause-and-effect relationships. Rather, it can be expected that there are complex interlinkages of multiple controls that favor hydrogeological similarity in different manifestations of the individual attributes.

We, therefore, confirm our initial expectation that similar GWL dynamics derived from indices are associated with multiple processes and, vice versa, that the important controls and processes can be identified and distinguished by a combination of indices. This is in line with previous findings with hydro-climatic time series (McMillan 2020; Beck et al. 2016).

The results show that relationships between GWL dynamics patterns and controls can be generalized to the global scale to some extent. Notably, the importance of different attribute categories (Table 2; Figure C12) for shaping GWL dynamics is related to hierarchical relationships where climate, topography, and soil conditions have a greater influence over the occurrence and characteristics of surface waters, as well as of land use and land cover, than the other way around. This implies that the direct influencing factors offer more explanatory power. Then again, direct influencing factors in the attribute dataset are more often based on actual observations and often have a higher resolution (e.g., the minor explanatory power of attributes from surface waters might also be related to a coarse spatial resolution of about 55 km). From the importance of elevation and water table depth, especially for the determination of CL3-type dynamics and differentiation from other dynamics that are more often found in near-surface aquifers, it can be concluded that these attributes mainly describe the increasing damping effect of groundwater recharge with increasingly long soil passages. The generally high importance of climatic attributes for GWL dynamics -– either only in the long term or also in short periods — is scientific consensus, making climate change a usual part of assessments of changes in groundwater systems (Riedel and Weber 2020). Also, within the group of climatic attributes, the importance of direct attributes such as temperature outweighs derived attributes such as the aridity index. According to Table 3, certain types of unconsolidated sediments and a terrestrial ecosystem can distinctly indicate or exclude specific GWL dynamic types. These manifestations of categorical attributes are likely to serve as proxies for the general importance of their class. Due to the multitude of manifestations, only the most common ones are likely to appear in the importance ranking in Figure C13 (e.g., eolian deposits mainly exist in the Netherlands, which also provides the most wells in the dataset).

Anthropogenic activities such as irrigation and groundwater pumping typically have a localized effect. They have the potential to overprint different types of GWL dynamics even when the hydrogeological setting is the same (Sorensen et al. 2021). Together with small-scale lithological and hydrogeological peculiarities, they likely contribute the most to the variability of GWL dynamic patterns locally and regionally. Section 4.5.1 emphasizes the importance of the data basis, including its current limitations, for assessing

certain processes. However, from the comparably minor importance of anthropogenic attributes for GWL dynamics, it could be assumed that global groundwater quantity is less influenced by anthropogenic than by natural controls. On the one hand, this was also the expectation with using data derived from monitoring networks and after quality control. On the other hand, it is unclear how much of the explanatory content of the anthropogenic attributes is based on anthropogenic activities and how much is due to correlations with natural controls (Figure C11). Such interactions could be that irrigation is higher in dry climates and that there are more sealed surfaces where population density is high. Therefore, anthropogenic impact is also difficult to disentangle from natural controls in this study.

Moderate accuracy of the RF model means that parts of the GWL dynamics derived from indices are not well predicted based on the used attributes. Hence, machine learning tools that are trained to find cause-and-effect relationships face the same challenge of process-based groundwater models: limited availability and quality of explanatory data, especially subsurface information. Even if there is great progress in the availability and quality of geoscientific data, surface attributes derived for the global scale are still biased by uncertainty in the data itself (Peng et al. 2017), by coarse spatial resolution, and by the period they cover. In the presented approach, no dynamic attributes were used because they are not available for all attribute categories at the global scale. For instance, land use has a large influence on the water balance with land use conversion (Mishra et al. 2014). The fact that explanatory information is missing in the RF model for differentiating GWL dynamics at the local to regional scale can be observed, for example, in the case study region in northern Germany (Figure 12). Furthermore, as explained in Section 4.3.4, studies of global scope, such as this one, face limitations in accurately defining subsurface catchments.

4.5.3 Case study

Finally, the question also arises as to what extent it is possible to explain the generalized GWL dynamics derived with the index-clustering approach of this study regionally if attributes were only good enough. The inclusion of regional context using the case study from northern Germany (Figure 12) serves a dual purpose in our study. First, in Section 4.5.1, we assessed the robustness and validity of our proposed approach for identifying GWL dynamics at different scales. In this section, the case study serves as a critical test bed for understanding the chances and limitations of extrapolating explanations for patterns of GWL dynamics from the global to the regional scale. This case

study was selected because a hydrogeological map and related hydrogeological information exist (LfU-SH 2003), the GWL time series source dataset includes information on the aquifer storey of the wells, the well density is high, and the accuracy of the RF model is about the same in this region as it is globally. These conditions allow the study of spatial

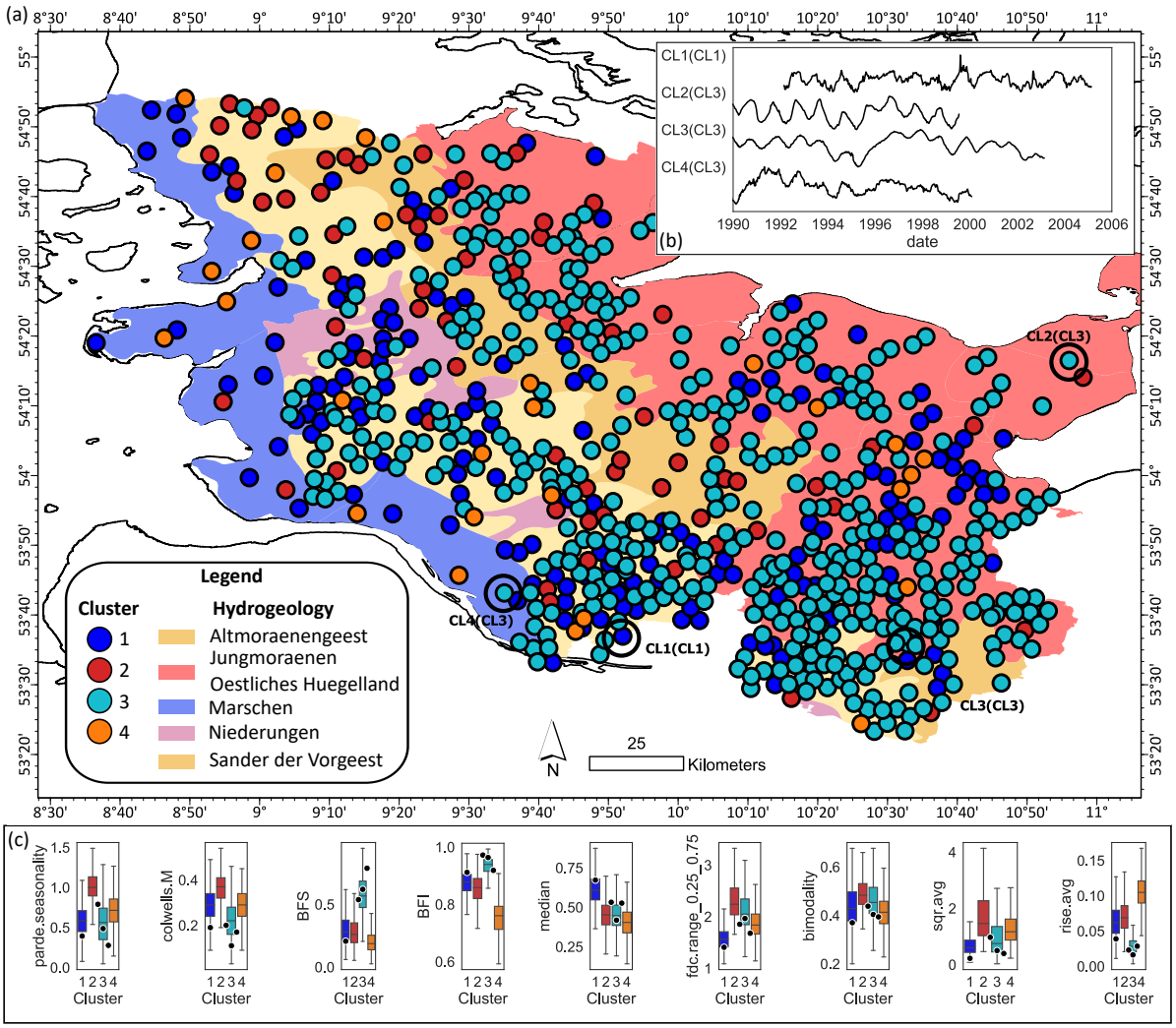


Figure 12: GWL dynamic patterns in the German federal state Schleswig-Holstein. (a) Cluster assignment to well locations and hydrogeology map of the German federal state Schleswig-Holstein (LfU-SH 2003). Overlapping well markers are jittered at a minimum spacing of 500 m and thus no longer represent the original well locations. The cluster number in parentheses marks the true cluster from k-means clustering. (b) Exemplary time series of each cluster from the encircled well sites in (a). (c) Within-cluster variability of the example indices that each represent a component of the Heudorfer et al. (2019) typology from Figure 10, plotted with index values (dots) of the encircled well sites in (a) where the dots between CL2 and CL3 and between CL4 and CL3 mark the wells that have the highest probability for CL2 and CL4 in the RF model but belong to CL3 according to the clustering (see Figure C8 for within-cluster variability for all indices that are plotted together with index values of the selected well sites).

patterns and the ground-truthing of the GWL dynamic patterns using expert knowledge.

To a large extent, the near-surface aquifer in this area is a pore aquifer of silicate type formed during the Ice Age (LfU-SH 2003; Otto 2001). The hydrogeological situation of the shallow groundwater can broadly be divided from west to east into three major landscapes: (a) low-lying marshlands (Marschen); (b) slightly raised Geest landscapes (Altmoraenengeest), the lower outwash plain (Sander der Vorgeest), and intervening depression zones comprising the river valleys (Niederungen); and (c) morainic uplands with stronger relief (Jungmoraenen Oestliches Huegelland).

GWL dynamics in the marshlands predominantly belong to CL1 (59%), especially in the upper two aquifer storeys (68%) and close to the sea, followed by CL3 (20%), which combines more GWL dynamics of the lower aquifer storeys (60%) and well sites at the border of the Geest. The Geest and uplands are characterized by GWL dynamics that mainly belong to CL3, with some exceptions, including the Geest landscape in the north close to the Danish border (GWL dynamics are of mixed type), the area around the major depression in the north (GWL dynamics mainly belong to CL1), and the eastern uplands at the southernmost bay (GWL dynamics mainly belong to CL1). In the lower Geest, GWL dynamics belong to CL2 (23%) more often than in other landscapes. As in the marshlands, CL3-type GWL dynamics are associated more often with lower aquifer storeys compared to the other clusters in these areas (63–78%), with the difference being that CL3 is also present to a similar degree in the upper aquifer storeys in the Geest and uplands (45–67%). Figure 12b displays example time series of wells that the RF model identifies as having the highest probability for each cluster. The cluster in parentheses indicates the actual cluster that is based, among others, on the index values shown in Figure 12c. In cases where CL3 was mistaken for CL2 and CL4, the example wells mostly have indices within the CL3 distribution (Figure 12c). While there are minimal visual differences in the time series between CL2 and CL3 at the well location where CL2 is predicted, some distinctions can be observed in Figure 12b when CL3 is misidentified as CL4. The inability of the RF model to correctly predict these instances can therefore mainly be attributed to two factors. First, the GWL dynamics at the predicted wells for CL2 and CL4 reside closer to the border between different dynamic types, making them more challenging to differentiate. This is confirmed by some very similar index values compared to CL3. Second, the limited explanatory potential provided by the attributes can be assumed to also contribute to the model's failure: expert information suggests that small-scale differences in GWL dynamics in the case study region can be explained by lower aquifer storeys with good hydraulic connections to upper storeys and surface waters. The presence of both CL1 and CL3 dynamics in the higher Geest and upland

regions, regardless of groundwater storey, can be attributed to differences in the soil covering, water table depth, and the existence of both unconfined and confined upper aquifers in these landscapes (LfU-SH 2003). Comparing the ground-truthing with the RF model, some global explanations apply to this regional context as well. For instance, in the marshlands and the northern depression zone, where impermeable sediments of small to medium thickness overlay the aquifer (LfU-SH 2003), the prevalence of CL1 and CL4 type wells is higher (Table 3).

4.6 Conclusions

In summary, this study provides new insights into the hydrogeological behavior and the hydrogeological similarity in coastal regions. It utilizes an unprecedented dataset of GWL time series and demonstrates the development of a global information archive on GWL dynamics for high-level generalizations. These findings can aid in developing broader frameworks for effective groundwater management – for example, by identifying regions with common characteristics and highlighting essential hydrograph features for global modeling. Our results show that hydrogeological similarity is a global phenomenon that, with respect to the aquifers near the coastline, can be divided into four clusters representing distinct patterns of GWL dynamics. Similar clusters are observed globally across different continents and climate zones, while GWL dynamics can be highly variable regionally and locally, suggesting a complex interplay of controlling factors. Therefore, extrapolating GWL dynamics from a single well to the surrounding area may sometimes be less appropriate than inferring GWL dynamics from global relationships. However, establishing global dynamics–control relationships remains a challenge. We found that these are mainly dominated by (a) topography and groundwater depth, mainly determining the responsiveness of groundwater systems to impacts on the water cycle that separates well a single cluster from three others, and (b) climate and soil characteristics, which differentiate these three clusters with high short-term and interannual variability. However, groundwater management typically requires a more precise explanation of GWL dynamics by controlling factors beyond what is currently achievable with the attribute data available at the global scale. In particular, disentangling anthropogenic impacts from natural controls in GWL dynamics, e.g., to accurately predict the effects of climate change on GWLs without incorrectly accounting for water withdrawals, is of great importance for sustainable groundwater management. While the overall importance of anthropogenic activities to GWL dynamics was found to be smaller compared to natural characteristics in this study, more in-depth analyses were beyond the scope of this study. When analyzing certain processes, including SWI, it is important to pay attention to the data basis

(selection of groundwater data and indices). To address the explanatory limitations of data-driven analysis and the prediction of GWL dynamics, we suggest using attributes closely related to actual observations and the direct influencing factors in combination with explainable machine learning techniques and suggest relying more on groundwater hydrographs and derived indices when attribute data are scarce or associated with large uncertainties. Overall, machine learning techniques combined with hydrograph information show promise for improving our understanding and predictive capabilities in addressing GWL complexities, especially at large spatial scales.

5 Multi-site deep learning for groundwater level prediction across global datasets: toward scalable applications under data scarcity

5.1 Abstract

Deep learning, and especially long short-term memory (LSTM) networks, have shown strong potential for GWL prediction, but remain limited across sites and at scale due to data scarcity and reliance on sparse, inconsistent, or physically ambiguous static site descriptors. This study advances prediction in three key ways. First, we present the first global, large-sample evaluation of multi-site LSTM models trained on over 1,800 wells across nine global regions and evaluate their performance under realistic data constraints. Second, we show that trainable site embeddings enable the model to learn site-specific behavior directly from time series without relying on externally defined site descriptors. Third, we analyze the embedding space, revealing emergent spatial and functional patterns that reflect hydrogeological structure. Embedding-based models achieve strong predictive performance in data-rich regions (median Nash-Sutcliffe efficiency (NSE) > 0.7) and remain robust across a combined global dataset. Our findings further challenge the assumption that larger datasets naturally improve predictions, especially in data-sparse regions. These findings describe both the potential and current limitations of scalable, entity-aware LSTM models for multi-site GWL prediction under data scarcity, paving the way for truly geographically global deep learning in hydrology.

5.2 Introduction

Groundwater supplies water for half of the world's domestic use and 38% of the world's irrigated land (United Nations 2022). As part of the hydrological cycle, groundwater influences climate regulation and supports biodiversity and ecosystem services (Gleeson et al. 2020b). It requires observational data and modeling techniques to gain an understanding of past and future trends in groundwater recharge and availability (Sanford 2002; Scanlon et al. 2002). Currently, in situ monitoring provides the most accurate insights into groundwater availability, with GWL data being the most commonly available parameter, enabling groundwater modeling and trend estimations. However, GWL data often remains sparse in many places (Lall et al. 2020; United Nations 2022), and GWL prediction is further complicated by the inherent complexity of its dynamics and the

numerous components and interactions within groundwater systems, driven by both climate and human activities (de Graaf et al. 2019; Gleeson et al. 2020a).

Traditional numerical groundwater model approaches approximate or simplify system complexity. Data scarcity, as well as the time-consuming data standardization and integration process, often limit these approaches. In addition, addressing the potential impacts of climate change on future groundwater resources generally requires high-quality, long-term, continuous GWL time series, which are rarely available in many regions worldwide. Therefore, GWLs are often analyzed with a limited number of representative wells for large regions (e.g., Bloomfield et al. (2019) and Cuthbert et al. (2019b). This underscores the necessity of bridging the gap between the consideration of localized hydro(geo)logic processes based on local data and large-scale effects, a critical step for water management in practice (United Nations 2022). It calls for innovative approaches that can effectively scale predictions by integrating diverse datasets and capturing complex interactions across different spatial scales (Massei et al. 2020; Gleeson et al. 2021). In this regard, data-driven algorithms have significantly advanced, with deep learning (DL), a subset of machine learning (ML) characterized by its use of layered neural networks, emerging as a particularly promising tool for accurate GWL prediction with minimal a priori process understanding at well locations (Rajaee et al. 2019; Tao et al. 2022). Beyond groundwater, DL and other ML approaches have been widely applied in hydrology, including rainfall forecasting (Mekanik et al. 2013), streamflow and runoff prediction (Kratzert et al. 2019; Balacumaresan et al. 2024), water supply and demand modeling (Mohammadi et al. 2024), and water quality monitoring (Najah et al. 2013), demonstrating their broad potential for capturing nonlinear hydrological processes across domains (Shen and Lawson 2021).

Early DL efforts in surface and subsurface hydrology were primarily focused on training models for individual wells with carefully selected high-quality time series (Lee et al. 2019; Solgi et al. 2021; Wunsch et al. 2022; Chidepudi et al. 2023). More recently, attention has shifted toward multi-site DL architectures – specifically Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber 1997) – trained on large, spatially distributed datasets, enabling simultaneous simulation of hydrological behavior across multiple locations. In this context, neural networks are trained on time series from numerous spatially distributed sites, allowing them to learn shared patterns while retaining local specificity. A key methodological advancement enabling this transition has been the introduction of entity-aware modeling by Kratzert et al. (2018) and Kratzert et al. (2019), which equips models with the ability to distinguish between individual sites, or "entities," during training and prediction. This is typically achieved by providing the

models with two types of input: dynamic input features such as meteorological forcing and static features relevant to describing the physical environment (e.g., hydrogeological and topographic attributes). These static features are intended to inform the model about inter-site (dis)similarities in hydrological functioning, thereby guiding the learning of inter-site similarities and improving predictive skill across sites (Kratzert et al. 2019; Heudorfer et al. 2024; Kunz et al. 2025).

In surface hydrology, standardized datasets like CAMELS (Addor et al. 2017), provide well-tested, spatially aligned static features that support this modeling paradigm for large spatial scales. In hydrogeology, however, no comparable global or continental-scale datasets exist. Groundwater systems are characterized by high spatial heterogeneity, limited observability, and often incomplete or uncertain information regarding aquifer properties, geology, or anthropogenic impacts. Furthermore, the derivation of static features frequently relies on proximity-based proxies (e.g., using nearby or regional land cover, soil, or topography to represent local hydrogeological properties) or aggregated statistics within spatial buffer zones around wells to capture environmental characteristics, which can introduce additional uncertainty due to spatial heterogeneity and limited representativity of these features for subsurface dynamics (Haaf et al. 2023; Heudorfer et al. 2024; Kunz et al. 2025). Recent studies have begun to question the added value of such static features in LSTM-based hydrological models. Evidence from both groundwater and surface water applications shows that time-series-derived statistics, random site identifiers, or one-hot encoded site labels can yield predictive performance comparable to that of models using detailed environmental descriptors, implying that static features often function more as unique entity identifiers than as meaningful representations of hydrological or hydrogeological processes (Li et al. 2022; Heudorfer et al. 2024; Heudorfer et al. 2025). Under this assumption, their absence may not substantially reduce model performance, raising the question of whether models can instead learn site-specific behavior and shared representations directly from time series.

In this study, we follow this assumption and investigate whether LSTM-based multi-site GWL prediction models can learn site-specific behavior and potentially shared representations directly from dynamic inputs and site identifiers, without relying on predefined static descriptors. Our approach is based on modeling and evaluating GWL time series from over 1,800 wells across five continents, encompassing a wide range of data quality, record lengths, and sampling frequencies. In contrast to prior studies using high-quality or pre-screened time series (e.g., Heudorfer et al. 2024; Chidepudi et al. 2025), we evaluate the performance of the model under realistic global conditions. This includes time series of varying completeness, noise levels, and hydrological contexts, which

are conditions commonly encountered in practice but rarely modeled. In doing so, we address a critical question: whether multi-site LSTM models can offer accurate, scalable solutions for groundwater prediction under data scarcity conditions – both in terms of time series and static feature availability.

As part of our investigation, we implement embeddings as a form of self-learned entity identification, assigning each monitoring well a trainable numerical vector that is updated during model training. This enables the model to capture site-specific behavior and inter-site similarity without requiring explicit physical metadata – unlike one-hot encodings, which treat all sites as unrelated (e.g., Chidepudi et al. 2025). The approach is comparable to the process how embeddings in natural language processing capture contextual relationships between words (Akbik et al. 2018; Wang et al. 2020). A similar mechanism was applied by Clark et al. (2022) in a multi-site DeepAR model for GWL prediction, where categorical embeddings were used to represent individual wells. However, the structure and meaning of the learned embeddings were not analyzed or interpreted by the authors. In contrast, our study introduces an explicit, trainable embedding layer and evaluates its utility for site differentiation and the internal representation of cross-site variability under data-scarce conditions. This assessment is inherently constrained by the same lack of independent physical descriptors that initially motivated the use of embeddings over predefined static features. Recent work in surface hydrology has demonstrated that learned internal representations in DL models can capture meaningful hydrological behavior. Lees et al. (2021) showed that LSTM cell states can reflect soil moisture dynamics, while Bassi et al. (2024) used autoencoder-derived landscape fingerprints to explain streamflow variability. Botterill and McMillan (2023) developed convolutional encoders to derive low-dimensional hydrological "signatures" from time series, offering an interpretable alternative to classical catchment descriptors.

The rest of this paper is organized as follows. Section 5.3 introduces the data and well locations. Section 5.4 outlines our methodology, employing LSTMs in an explorative approach, where we use this model architecture with either only dynamic input data or with additional site identifiers provided through one-hot encoding or trainable embeddings. Section 5.5 presents the results and discussion, focusing on both model performance and the explanation of embeddings to address our study objectives and their broader implications. Finally, Section 5.6 concludes the study by outlining its contributions to advancing large-scale environmental modeling using DL, particularly in the context of data scarcity.

5.3 Study area and data

Precipitation and temperature data were used as dynamic input features and originated from the ERA5 dataset, the fifth generation of atmospheric reanalysis by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hersbach et al. 2020). Additionally, we used a third dynamic input feature that has proven valuable for GWL prediction in previous studies (Wunsch et al. 2021a; Heudorfer et al. 2024). This feature involved fitting an annual sinusoidal curve to the temperature data, which potentially allows the model to capture seasonal patterns more effectively.

The modeled GWL time series originate from a global dataset previously compiled to investigate coastal GWL dynamics within 100 km of the shoreline (Nolte et al. 2024). The corresponding data sources are listed in Table 1. For this study, we selected shallow wells with a maximum median water table depth of 20 m to focus on near-surface conditions. The CoastalDEM elevation dataset (Kulp and Strauss 2019) was used to convert groundwater elevation data to GWLs referenced to the ground surface (if applicable), ensuring a consistent reference point across the dataset. Wells in semi-confined or confined aquifers were excluded when such information was available, as these aquifers often show different hydraulic behaviors compared with unconfined aquifers. This makes their dynamic features less comparable. From this data pool, GWL time series were selected and preprocessed. We chose GWL data from the period 1979 to 2018, because this period aligns with the ERA5 climate data availability. The time series were aggregated to a weekly resolution and clustered with respect to geographic regions, creating nine datasets with over 200 GWL time series each.

Table 4: Criteria for selecting GWL time series per focus region (Figure 13a), including length, availability, and gap length; median GWL per focus region. More information on the selection criteria is available in Section 5.3.

Criteria	**AU**	**DE**	**FR**	**IN**	**NA-N**	**NA-S**	**NL**	**Scand**	**ZA**
Length [years]	9	15	6*	15	12	12	12*	15	4
Availability [%]	40	90	90	7	20	20	90	20	20
Gap length [weeks]	12	2	2	26	12	12	2	12	12
GWL median	-5.0	-4.5	-5.0	-4.6	-3.8	-3.2	-1.2	-2.9	-4.5

* Initially, about 30 GWL time series with a minimum length of 15 years were chosen. Afterward, the remaining GWL time series that met the length criteria given in the table were randomly selected from the dataset.

The following selection criteria were applied, with specific values provided in Table 4: (1) the time series length in years (length); (2) the proportion of the data points in weekly aggregated time series that have GWL data available (availability); and (3) the maximum duration, measured in weeks, during which no GWL data is available (gap

length). In some cases, the data selection includes only segments of the time series, i.e., not necessarily the complete time series. Time series selected using these criteria then underwent visual screening – a step still recommended to identify and minimize errors, noise, and anthropogenic effects inside the GWL data (Barthel et al. 2022; Retike et al. 2022). We found data points significantly outside the value range (outliers), sudden sharp level changes over specific periods, and patterns potentially indicative of well pumping or anthropogenic recharge events, guided by our own experience and illustrative examples shown in Retike et al. (2022). Additionally, we applied Density-Based Spatial Clustering of Applications with Noise (DBSCAN; Ester et al. 1996) to rescue a few GWL time series by grouping similar data points into clusters based on their spatial and temporal proximity while identifying and excluding outliers. Finally, we randomly selected 200 time series for each refined regional dataset to ensure equally large sample sizes for every region and enhance comparability. These 200 wells were drawn from initial regional pools ranging from approximately 280 to 480 wells each. Geographical delineations of focus regions are illustrated in Figure 13a. Based on the nine focus regions (Figure 13b), the full, global dataset includes 1,800 GWL time series that vary in data quantity, including periods covered, data gaps, and overall sample sizes (Figure D1). The focus regions also span a wide range of climatic conditions (Figure D2). Rather than being perfectly physiographically representative or reflective of typical data quality and quantity in their respective regions, the datasets reflect the diverse and often inconsistent nature of global groundwater datasets.

For the modeling task, each time series dataset at the selected well locations (GWL, precipitation, temperature, and temperature seasonality) was standardized to have a mean of zero and a standard deviation of one, addressing discrepancies in scale and the impact of errors across different time series. This standardization, which was fitted on the training portion of the data (Section 5.5.2), ensures that no single series dominates the learning process.

5.4 Methods

5.4.1 Scope and experimental design

This study focuses on model development for GWL prediction by evaluating the utility of LSTM-based models in data-scarce, large-scale settings, where both time series and static input features are limited. While benchmarking against process-based, hybrid, or alternative ML models remains an active area of research, previous studies have demonstrated

that LSTMs often outperform (i.e., achieve higher predictive accuracy, such as improved Nash-Sutcliffe efficiency (NSE) or lower RMSE when capturing observed groundwater or streamflow dynamics) or complement traditional approaches for hydrological time series prediction and are widely regarded as a state-of-the-art, well-established DL method in this context (Kratzert et al. 2019; Rajaee et al. 2019; Lees et al. 2021). Building on these findings, we assess LSTM performance under real-world data constraints and investigate the role of trainable embeddings for representing site-specific behavior.

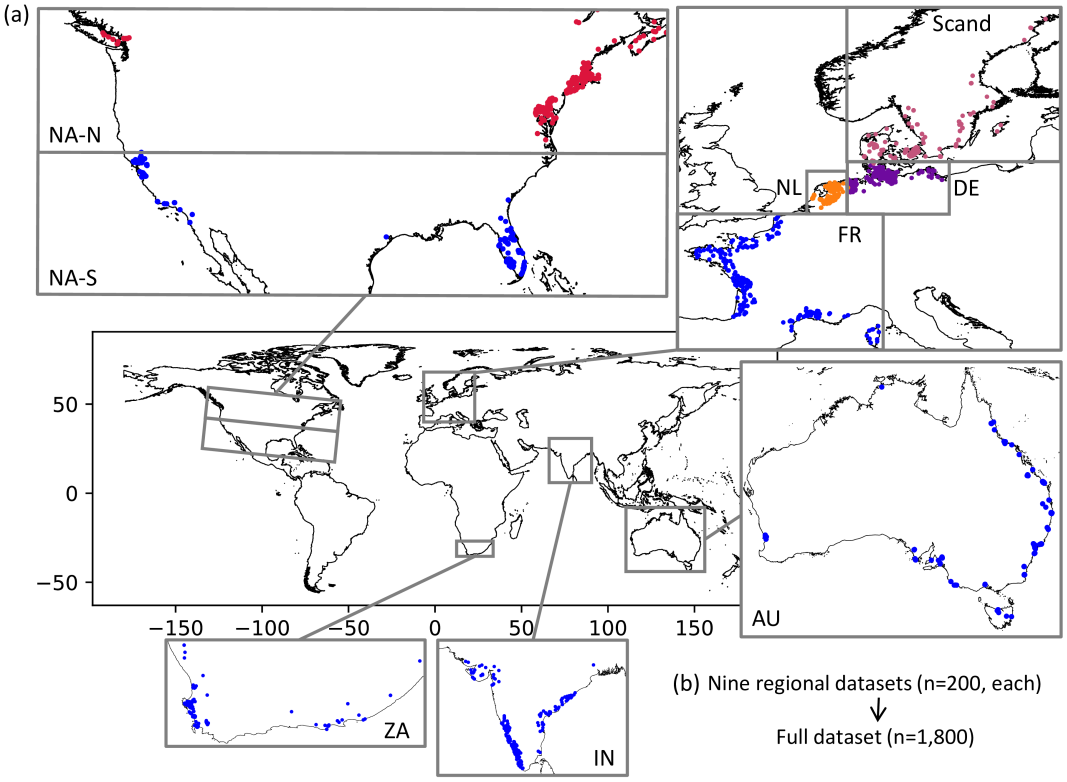


Figure 13: (a) Maps of the focus regions highlighting the geographical areas from which regional datasets have been sourced for this study: Australia (AU); Germany (DE); France (FR); India (IN); North America – North (NA-N); North America – South (NA-S); Netherlands (NL); Scandinavia containing Denmark and Sweden (Scand); South Africa (ZA). Colored dots represent the well locations of selected GWL time series. (b) Overview of model application scale. Each of the nine regional datasets contains 200 wells, collectively forming the full, global dataset used in the global modeling experiments (n = 1,800).

We tested the prediction capabilities of LSTM models within a multi-site architecture in a series of experiments using two setups (Figure 14): a standard configuration (Section 5.4.2) and a modified setup with embeddings (Section 5.4.3). Both model setups were applied to each of the individual regional datasets as well as to the global dataset (Figure 13b). In the standard setup, we used either exclusively dynamic input data comprising precipitation

and temperature (*nofeat* models), or static site information in the form of one-hot-encoded well identifiers (ohe features; *ohefeat* models), which were concatenated to the dynamic inputs at each time step. This allows the multi-site model to distinguish between different groundwater dynamics of individual wells. Similarly, the embedding-based setup (*embeddings* models) uses well identifiers to generate trainable vectors that serve as compact representations of individual wells. In this setup, the matrix from the static input has the dimension of the embeddings instead of being multiplied by the number of wells as in the *ohefeat* models. To explore how embeddings might enhance model interpretability and capture spatially meaningful patterns, we applied dimensionality reduction and clustering techniques (Section 5.4.3). We structured the evaluation of our modeling framework around three main analytical dimensions:

1. **Model architecture comparison (Sections 5.5.1 and 5.5.2):** We compared LSTM models with different site representations (no static input, one-hot encoding, and embeddings) across all regional datasets. This allowed us to assess the role of dynamic and site-specific information and evaluate the benefit of embedding-based representations.

2. **Training data scope, quantity, and cross-regional learning (Sections 5.5.3 and 5.5.4):** We evaluated the effects of data volume, diversity, and completeness by comparing regionally trained models among themselves and against a global model, assessing trade-offs between specialization and the integration of diverse datasets under varying data conditions.

3. **Learned spatial representations and internal differentiation across sites (Section 5.5.5):** We analyzed the embedding space to assess whether site-specific groundwater dynamics were internally encoded in a way that reflects spatial and functional similarity, enabling differentiation across wells without predefined static features.

5.4.2 Standard LSTM

We applied a standard LSTM architecture in which all input features enter the model through the same input layer (Hochreiter and Schmidhuber 1997). For more details on the model architecture, we refer to previous applications, such as those by Kratzert et al. (2019). The LSTM model was trained to make predictions at a weekly resolution.

In the multi-site architecture, data from multiple wells were combined: all time series were aligned chronologically, and missing values were retained as gaps. Each GWL time series was split into three set parts for training (70%), validation (15%), and testing (15%). We then pooled the sequences from all wells into three "buckets" (train/stop/test). We conducted a preliminary investigation on the validation set to assess general performance and examine the effects of the newly implemented embeddings (Section 5.4.3).

Model performance was primarily assessed using the Nash-Sutcliffe efficiency (NSE), calculated on the independent test set to ensure unbiased evaluation. High NSE values (closer to 1) indicate strong agreement between model predictions and observed variance during the test period. In addition to NSE, we computed several complementary metrics: Kling-Gupta efficiency (KGE), coefficient of determination (R^2), bias, mean squared error (MSE), and RMSE. However, NSE was used as the primary performance indicator throughout the analysis.

LSTM networks are designed to capture long-term dependencies in time series data through memory cells and gating mechanisms, making them well-suited for irregular temporal patterns. However, they do not inherently handle missing values, which must be explicitly addressed during training. We applied a masking technique that excluded missing values from error calculations, ensuring that training and evaluation focused only

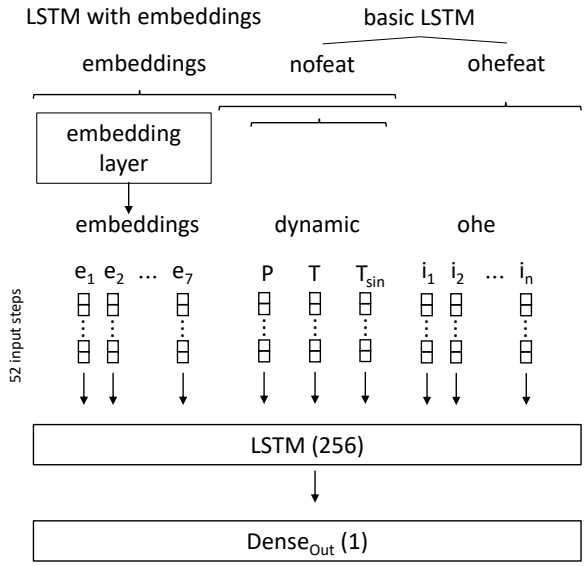


Figure 14: Long Short-Term Memory (LSTM) model structure for the three variants: *nofeat* (three dynamic input features only), *ohefeat* (dynamic inputs combined with one-hot encoded well identifiers i, increasing the input dimensionality by n, the number of individual GWL time series), and *embeddings* (dynamic input features combined with seven learned well embeddings e).

on available data points.

The MSE served as the loss function in the training process. During training, the Adam optimizer (Kingma and Ba 2014) was used to find the set of weights that minimizes this loss. The standardization of the GWL time series (see Section 5.3) prevented specific wells, particularly those with larger fluctuations or inherently larger values, from disproportionately influencing the loss. Standardization also aligned the scale of the MSE used in training with that of the NSE used in testing, ensuring that error metrics were consistent and comparable across wells in a multi-site setup (Heudorfer et al. 2024).

Hyperparameters control both the architecture and the learning procedure of the LSTM. Following common practice in large-sample hydrology, we used a single-layer LSTM with 256 hidden units, followed by dropout and a dense output layer (Figure 14; full settings in Appendix D.3). This architecture was originally selected in Kratzert et al. (2019) via grid search with cross-validation on the CAMELS basins, has therefore been applied across diverse hydroclimatic settings, and is now a common baseline (e.g., Frame et al. 2022; Klotz et al. 2022; Loritz et al. 2024; Heudorfer et al. 2025). To keep model capacity appropriate for dataset size and aligned with prior applications of this architecture (typically mid-hundreds of basins/wells), we limited each regional dataset to 200 wells so that regional and global experiments were on a comparable scale. We did not perform additional hyperparameter optimization to maximize comparability across setups and datasets, and because we aimed to understand capabilities and limitations of the LSTM models rather than to push absolute performance. In this context, further tuning typically brings incremental improvements rather than categorical differences when the model is already reasonably sized, i.e., not excessively large or small.

To assess the robustness of the model results, each of the three model variants (Figure 14) was trained and evaluated 10 times using different random seeds, resulting in distinct weight and bias initializations. While more advanced uncertainty quantification methods (e.g., Monte Carlo dropout) exist (Klotz et al. 2022), this approach follows standard practice and is sufficient for model robustness evaluation in similar hydrological DL studies (e.g., Kratzert et al. 2019; Chidepudi et al. 2025; Mangukiya and Sharma 2025). It allowed us to evaluate performance variability across training runs and ensured that findings were not driven by a single initialization.

5.4.3 LSTM with embeddings

The model architecture in the LSTM with the *embeddings* model setup is similar to that in the standard LSTM model setup described in Section 5.4.2, with one exception: the substitution of the one-hot encoding with an embedding layer (Figure 14). Like the *ohefeat* models, the encoding process of the *embeddings* models differentiates between individual wells in the dataset. However, well identifiers are converted into a dense, lower-dimensional space in the embedding layer. Specifically, we employed a seven-dimensional embedding space for representing well identifiers in our LSTM models. This dimensionality was chosen to balance computational efficiency and the capability to capture and distinguish well characteristics across the large dataset. By doing so, it greatly reduces the number of input dimensions compared with one-hot encoding.

For each well identifier, an embedding is retrieved from an embedding matrix E, where each row corresponds to a unique well identifier. This can be expressed as:

$$E_i = E\Big[X_{\text{static}}[i]\Big]$$

where E_i represents the embedding for the ith well identifier in the input vector X_{static}. While these embeddings do not encode wells uniquely in the way one-hot encoded vectors do, they may help the model capture shared behavior patterns among wells based on their time series.

The embedding layer, implemented using the Python library Keras, learns an embedding matrix during training to optimize the representations of the individual wells for the task. The embeddings are adjusted during the training process to minimize the prediction error by capturing similarities in patterns within the time series. Specifically, the embeddings are optimized to predict GWLs using combined static (embedding-derived) and dynamic (meteorological) inputs. After this transformation, both inputs are concatenated and fed into the LSTM layer, similar to the *ohefeat* model variant (Figure 14).

5.4.4 Exploring learned representations through embeddings

The embedding layer (Section 5.4.3) provides an opportunity to investigate how the model organizes information about individual wells based on their time series. LSTM models are often regarded as "black boxes" due to their complex internal states and lack of direct interpretability. Although embeddings do not constitute explainability in the classical

sense of feature attribution (e.g., SHAP or LIME), they allow for post hoc analysis of how the model internally differentiates between sites.

To investigate these learned representations, we analyzed the distribution of the embeddings values in relation to geographical location and GWL dynamics, which have previously been linked to environmental descriptors such as topography and lithology (Nolte et al. 2024). Both linear (PCA) and nonlinear (t-distributed Stochastic Neighbor Embedding, t-SNE; van der Maaten and Hinton 2008) dimensionality reduction techniques were used to project the seven-dimensional embedding space into two dimensions for visual exploration. The resulting projections were then visualized using color-coding to assess whether the embeddings captured meaningful structure. Two coloring strategies were employed: (a) based on geographical information, to reflect large-scale climatic influences and (b) based on clusters of GWL dynamics previously defined by Nolte et al. (2024). For (a), wells were categorized by geographical regions (continents and countries). For (b), wells were assigned to one of four clusters derived from statistical patterns in hydrographs (Heudorfer et al. 2019) representing different types of GWL behavior (e.g., stable seasonal cycles, rapid recharge responses, and interannual variability). These clusters were based on approximately 8,000 globally distributed groundwater hydrographs and included nearly all wells from the DE dataset and about 35% of the wells in the full dataset of this study.

Beyond visual comparisons, a quantitative method can help to assess the alignment between the learned embeddings and the prior GWL dynamics clusters. Specifically, we compared the four prior clusters with four newly generated clusters obtained by applying k-means clustering to the embeddings (using the same settings and number of principal components as in the prior study). Cluster cohesion and separation was evaluated using silhouette scores (Rousseeuw 1987), where higher values indicate more internally consistent clusters. In addition, the normalized mutual information (NMI; Strehl and Ghosh 2002) was used to compare both sets of clusters, i.e., to quantify the correspondence between the learned and reference clusters. In addition, we used DBSCAN (eps = 0.25, min_samples = 2) to identify inliers and outliers in the embedding space, aiming to detect wells that were isolated or underrepresented in the learned representation. The subsequent comparison of model performance metrics between these groups aimed to assess whether sparsely represented wells were associated with lower predictive skill.

5.5 Results and discussion

5.5.1 Model performance across regions

This section presents the results of LSTM models trained and tested on regional datasets only. Across the nine regions, a total of 270 model runs were conducted, covering three model variants and accounting for variability through multiple training initializations. The resulting model performances, summarized in Table 5 and Figure 15, show how predictive skill varies by region and input configuration. Spatial patterns in model accuracy across individual wells are illustrated in Figure D3, showing regional and local variation within the datasets.

Table 5: Mean, 10th percentile (Q10), and 90th percentile (Q90) Nash-Sutcliffe efficiency (NSE) scores across 10 random initializations, evaluated for the three model variants *nofeat*, *embeddings*, and *ohefeat* (Figure 14) across the nine focus regions shown in Figure 13a.

NSE		**AU**			**DE**			**FR**	
	Q10	Q50	Q90	Q10	Q50	Q90	Q10	Q50	Q90
nofeat	0.425	0.495	0.544	0.399	0.462	0.529	0.549	0.594	0.626
embeddings	**0.670**	**0.701**	**0.739**	**0.708**	**0.729**	**0.746**	0.774	0.788	0.798
ohefeat	0.637	0.670	0.691	0.664	0.701	0.745	**0.781**	**0.795**	**0.806**
NSE		**IN**			**NA-N**			**NA-S**	
	Q10	Q50	Q90	Q10	Q50	Q90	Q10	Q50	Q90
nofeat	**0.489**	0.509	0.524	0.522	0.536	0.554	0.382	0.398	0.414
embeddings	0.488	**0.523**	**0.552**	**0.679**	**0.692**	**0.701**	**0.437**	**0.459**	**0.477**
ohefeat	0.448	0.466	0.489	0.674	0.682	0.690	0.381	0.404	0.421
NSE		**NL**			**Scand**			**ZA**	
	Q10	Q50	Q90	Q10	Q50	Q90	Q10	Q50	Q90
nofeat	0.424	0.488	0.562	0.362	0.385	0.415	0.369	0.398	0.440
embeddings	**0.674**	**0.688**	**0.708**	**0.569**	**0.585**	**0.609**	0.496	**0.599**	**0.667**
ohefeat	0.655	0.675	0.698	0.446	0.492	0.537	**0.509**	0.552	0.589

Additional evaluation metrics are available in Table D1. Bold values indicate the highest NSE score among the three model variants (*nofeat*, *embeddings*, *ohefeat*) for each region and quantile.

Models using embeddings and one-hot encoded site identifiers (*ohefeat*) consistently outperformed the variant without static input features (*nofeat*), as discussed in more detail in Section 5.5.2. The *embeddings* variant achieved the highest median performance in eight out of nine regions. In five regional datasets – AU, DE, FR, NA-N, and NL – these models achieved moderate to high performance, with median NSE scores ranging between 0.67 and 0.80. In contrast, moderate to low performance, with median NSE scores from 0.40 to 0.55, was derived for the datasets from IN, NA-S, Scand, and ZA. Regions with higher performance also exhibited lower variability in the models with site identifiers, as reflected by smaller interquartile ranges (Q90–Q10). For instance, in FR,

the *embeddings* and *ohefeat* models both achieved stable predictive skill, with NSE values ranging between 0.78 and 0.81. Additional evaluation metrics (Table D1) support these findings.

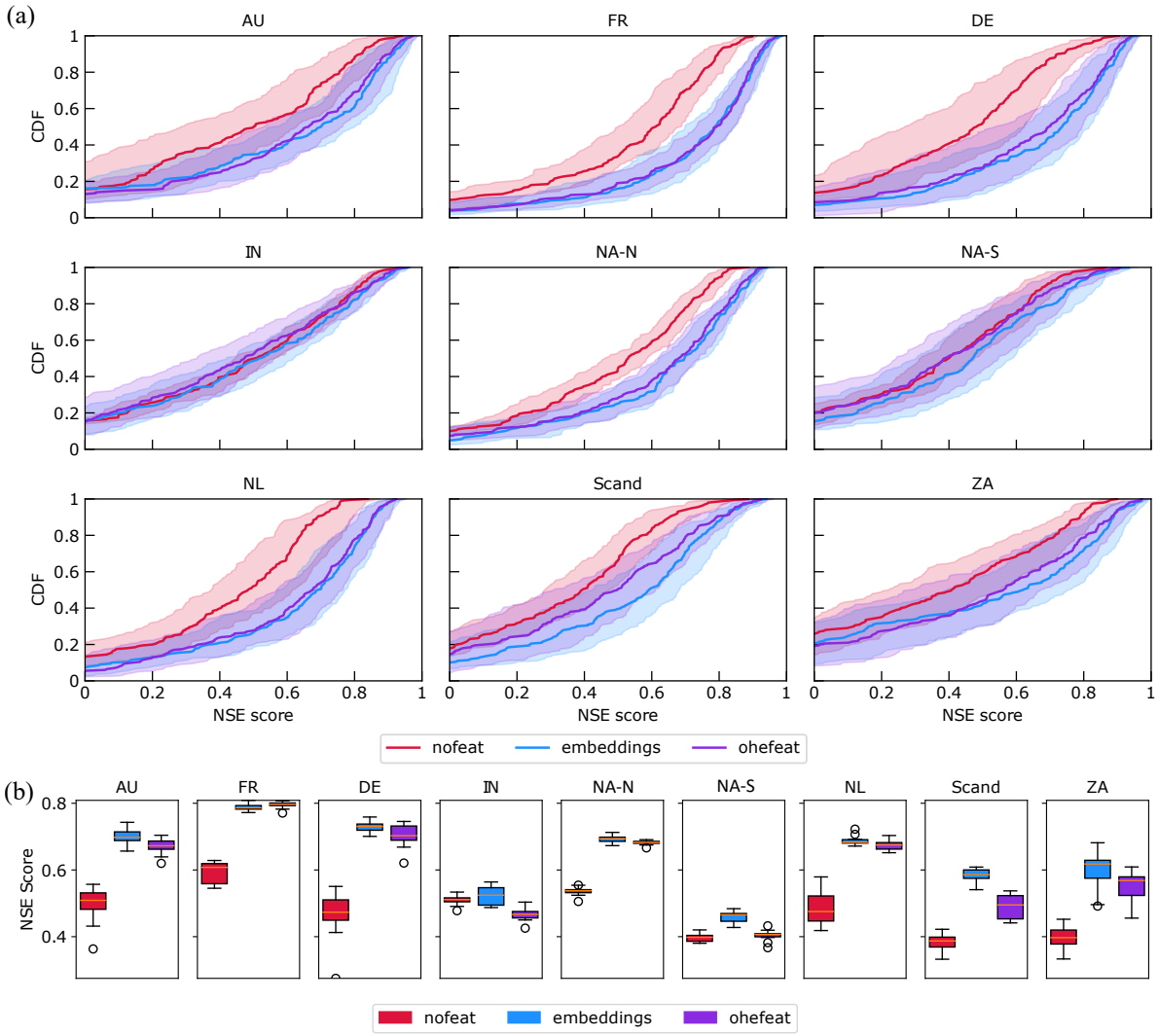


Figure 15: Nash-Sutcliffe efficiency (NSE) scores for the three model variants *nofeat*, *embeddings*, and *ohefeat* (Figure 14) across the nine focus regions shown in Figure 13a. (a) Cumulative Distribution Functions (CDFs) of individual well NSE scores. Solid lines represent the median across 10 models with different random initializations; shaded envelopes indicate the range of those runs. (b) Boxplots showing the distribution of median NSE scores from the 10 model runs for each region and model variant.

5.5.2 Superior performance of entity-aware models

While dynamic input features, which capture temporal patterns and enable learning from historical variability, are most important for GWL prediction, the inclusion of direct site identifiers in the *ohefeat* and *embeddings* models further improved performance by 16–37% (Table 5; Figure 15). Except for the datasets from IN and NA-S, the *nofeat* models show notably lower performance compared with the model variants with site identification. Among these, the *embeddings* models outperformed the *ohefeat* models across all datasets except FR, with median NSE improvements of up to 0.1 in several datasets.

Our findings of enhanced predictive capabilities in models with direct site identifiers align with previous studies that demonstrated the success of these models in learning hydrological patterns across catchments in the United States (Kratzert et al. 2019) and GWL dynamics in Germany (Heudorfer et al. 2024; Kunz et al. 2025). These models, often referred to as entity-aware, can help overcome current limitations posed by uncertainties in descriptive environmental data. However, we observed only marginal performance improvements from including site identifiers in the IN dataset. In this case, the GWL data are particularly sparse but regularly distributed (Figure D1), suggesting a potential upper limit to the ability of the model to discern patterns at the given temporal resolution and data quality (compare Section 5.5.4). Consequently, a minimum monthly data frequency appears necessary to support the effective use of site-specific static input.

Performance benefits with site identifiers are evident for most wells. The cumulative distribution functions (CDFs) in Figure 15a show a steeper ascent for the *embeddings* and *ohefeat* models compared with the *nofeat* model, indicating a higher proportion of wells achieving elevated NSE scores when site identification is included. Wells with medium to high baseline scores exhibit noticeable improvements in predictive accuracy. In contrast, wells with initially low scores show little to no performance gain, suggesting that modeling challenges in these cases are primarily driven by limited or compromised data. Similarly, a subset of wells with negative NSE values – indicating less accurate predictions than the average of the observed data – remain difficult to model across all model variants. This is consistent with previous multi-site LSTM studies in both surface and groundwater hydrology (Kratzert et al. 2019; Li et al. 2022; Heudorfer et al. 2024).

The *ohefeat* model requires substantially more computational resources due to increased input dimensionality, while the *embeddings* model provides a more resource-efficient alternative by reducing input space without compromising performance. This makes the embeddings approach particularly advantageous for large-sample datasets with extensive

GWL time series and for large-scale applications, such as the global dataset used in this study (see Section 5.5.3).

5.5.3 Global models match regional performance

To evaluate the effects of combining data from regions with richer, higher-quality GWL time series and those with less comprehensive or lower-quality data (compare Section 5.5.4), we conducted experiments using the entire global GWL dataset, which consists of 1,800 time series. In contrast to the regional model runs discussed in Sections 4.1 and 4.2 (Figure 15, Table 5), where models were trained separately for each region, the global model is trained on the complete, worldwide dataset, bringing together all available wells into a single training framework. Given the high computational demands of the *ohefeat* model, which become increasingly prohibitive as the number of time series increases, and the superior performance of the *embeddings* model (Section 5.4.3), we focused our comparison on the *embeddings* and *nofeat* models.

Figure 16 compares the performance of models trained on the full global dataset with those trained separately on each region. In Figure 16a, the CDFs of NSE scores show that both the global *embeddings* model and *nofeat* model perform almost identically to their respective ensembles of regional models. The median NSE scores from the global dataset are 0.44 (*nofeat* model) and 0.63 (*embeddings*), closely aligning with the median across all regional models combined (compare Section 5.5.1). This consistency also holds at the level of individual wells: as shown in Figure 16b, most points lie close to the 1:1 line, indicating that NSE values for individual wells remain largely stable when switching from regional to global training. These results have two key implications.

First, increasing the size of the training dataset introduces greater variability, which in theory can help DL models generalize better by exposing them to a broader range of patterns. However, this does not automatically improve model performance for region-specific or local GWL predictions. Previous studies have shown that LSTM models trained in multi-site setups often outperform single-site models (e.g., in Heudorfer et al. 2024). Furthermore, training on data-rich regions has been found to improve runoff estimations (Ma et al., 2021; Le et al., 2024) and GWL simulations (Clark et al. 2022) in data-sparse regions. However, unlike previous approaches that randomly combined data or grouped catchments based on hydrological similarity (Fang et al. 2022; Kratzert et al. 2024) or hydrogeological similarity to reduce data complexity (Chidepudi et al. 2025), our regional datasets were defined by geographic proximity. This means that each regional dataset contains a full array of time series from its respective geographic area, capturing a

wide range of GWL dynamics that can vary significantly. This variation often occurs even within small spatial extents due to subsurface heterogeneity and scale-dependent processes (Blöschl and Sivapalan 1995; Neuman and Di Federico 2003; Nolte et al. 2024). Therefore, the most relevant dynamics for each region (including both dominant and more localized temporal patterns) are likely already well represented within its dataset. We conclude that incorporating additional data from geographically and climatologically dissimilar regions (Figure D2) may not provide meaningful context for supporting the modeling task in this study. This is not because the model cannot integrate heterogeneous inputs, but because the added data do not contribute temporal or hydrogeological patterns beyond those already represented in the target region.

Second, while no additional benefits were found with the global model for the data-

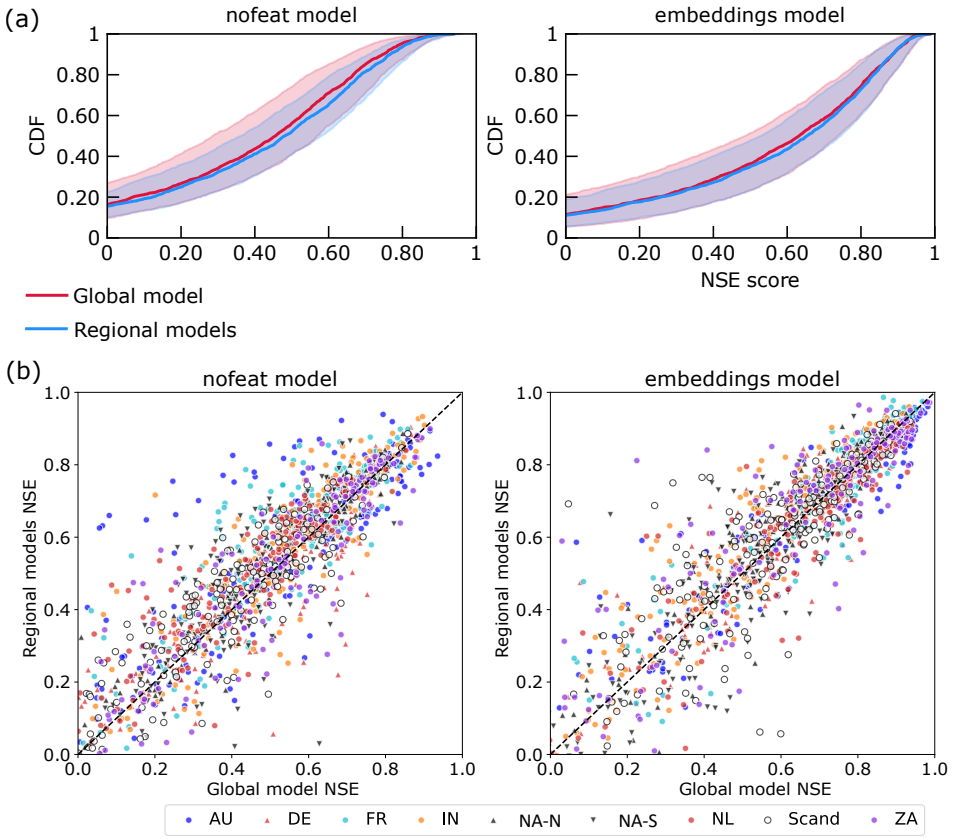


Figure 16: Comparison of Nash-Sutcliffe efficiency (NSE) scores for the *nofeat* and *embeddings* models trained on the global dataset (global model) versus models trained separately on regional datasets (regional models). (a) Cumulative Distribution Functions (CDFs) of individual well NSE scores. Solid lines represent the median across 10 model initializations; shaded envelopes indicate the range across runs. (b) Scatter plots comparing NSE scores from global models (x-axis) against regional models (y-axis) for individual wells, colored by focus region as defined in Figure 13a.

sparse regions in this study, the comparable performance of the global model to regional models highlights the potential for multi-site architectures to integrate diverse datasets from around the world without compromising prediction accuracy for individual regions or time series. The ability to effectively incorporate data from regions with varying hydrogeological conditions and data quality reduces the need for time-consuming time series selection. Furthermore, it underscores the potential scalability of these approaches with DL.

While the global model did not yield performance gains for data-sparse regions in this study, previous research has shown that comprehensive datasets can offer substantial benefits under different conditions. For example, transfer learning – where models are applied across regions (Wagenaar et al. 2018; Ma et al. 2021), typically supported by static input features that help guide cross-regional generalization – has demonstrated such benefits. In contrast, our setup omits such features, relying solely on dynamic inputs and learned site representations. Future experiments could build on this work by leveraging high-resolution datasets (e.g., daily) to improve model performance or simulate finer-scale dynamics in regions where only lower-resolution data (e.g., monthly) are available. Similar advantages have also been shown for addressing unforeseen climatic conditions not represented in the training data (Fang et al. 2022; Wi and Steinschneider 2024). Developing such capabilities is critical for advancing DL models toward practical deployment in operational GWL forecasting, though such experiments lie beyond the scope of this study. Overall, the multi-site framework presented here offers a promising foundation for these future directions by enabling modeling across diverse hydrogeological and climatic conditions – even at the global scale and without relying on predefined static features.

5.5.4 Data quantity and quality limitations

The observed differences in performance across regional models are closely linked to data quantity (the availability of GWL measurements within each time series; Figure D1). Figure 5 illustrates the differences by showcasing selected time series whose NSE scores are near the median for their respective region. In general, the model performance is much better in regions with greater data availability and fewer gaps (DE, FR, NL) compared with those with more limited or irregular data (IN, ZA). These findings further support that sufficient and relevant training data improve model performance – not only in single-well data-driven modeling (Wunsch et al. 2021b) but also in multi-site configurations. In LSTM models, low data availability or quality may hinder stabilization

during training, reducing predictive accuracy and increasing variability across runs. As discussed in Section 5.5.2 for the IN dataset, it appears that there are thresholds in data quantity below which the model struggles to learn and represent GWL dynamics effectively. These thresholds remain understudied in DL groundwater modeling, as most prior studies focused on high-quality, complete time series.

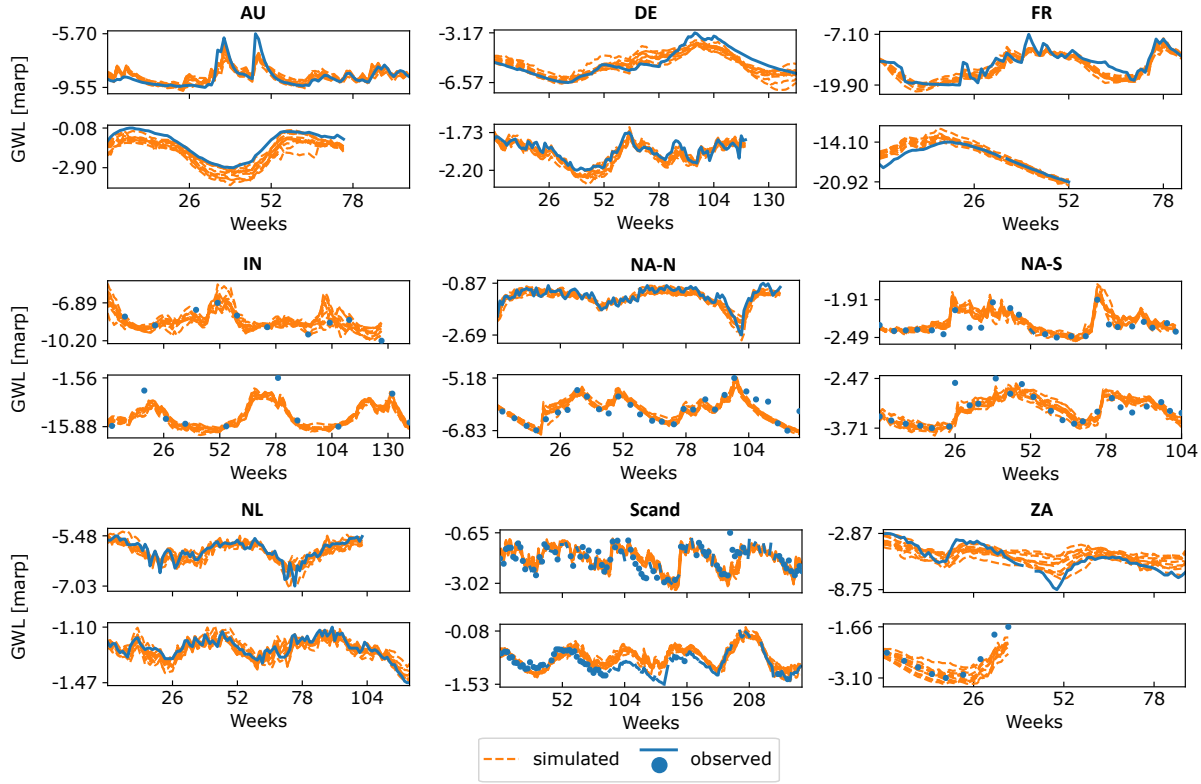


Figure 17: Example GWL time series in the test period from two wells in each of the nine focus regions (see Figure 13a), selected to approximate the median Nash-Sutcliffe efficiency value of their respective regional *embeddings* models. GWL is shown in meters above reference point (marp). Simulated values (orange dashed) are compared against observed measurements (blue). The figure illustrates regional differences in data resolution, and prediction accuracy.

Our results suggest that, beyond the general principle that more information creates better results, a minimum aggregation frequency of monthly observations appears necessary not only for reliable model performance but also for enabling the model to learn site-specific groundwater dynamics. This is evident from the IN models, which were the only ones among all regions to show minimal benefit from including site identifiers, likely due to the very low measurement frequency, with only a few observations per year. How aquifers respond to climatic inputs is difficult to capture with such sparse data. In addition, models in FR performed well despite being trained on fewer than seven years of data, suggesting that measurement frequency may be more critical for model performance than

the total length of the time series. High-frequency inputs allow models to better capture temporal variability, especially in systems influenced by strong seasonal or short-term fluctuations (Taylor and Alley 2001) and to avoid aliasing-effects (Bender 2007). Similarly, Wunsch et al. (2023) showed that high-frequency (weekly) inputs are essential for learning seasonal and short-term groundwater processes, which are often lost at coarser temporal resolutions.

By contrast, data with longer but consistent intervals smooth out variability, simplifying the prediction task. This effect is seen in the results for IN, where quarterly sampling intervals produced performance comparable to that of regions with irregular, sparse data (ZA, Scand). The irregular intervals and large gaps in these datasets make it more challenging for models to capture and adapt to inconsistent GWL dynamics – even though those dynamics might be equally or more predictable than in IN if the prediction targets were similarly aggregated (quarterly). In this sense, the seemingly better performance in IN may reflect a less complex prediction target, rather than inherently more predictable groundwater behavior. While time series length was less critical for our short-term, weekly-ahead predictions, it becomes essential for long-term groundwater forecasting. Climate-driven GWL trends often evolve slowly and are masked by low-frequency variability, requiring multi-decadal records to distinguish persistent changes (Baulon et al. 2022b), which long-term (DL) forecasts typically rely on (e.g., Wunsch et al. 2022). These aspects highlight the importance of aligning the prediction target with data properties when selecting GWL time series and evaluating model performance.

Other factors beyond data quantity that might influence model performance include the complexity of the system being modeled and the quality of the time series data. Characteristics of the modeled aquifer systems, such as groundwater depth, permeability, and aquifer type, can affect how aquifers respond to climatic inputs. While the regional datasets in this study are comparable in terms of groundwater depths (Table 4) and associated climate-groundwater interactions (Cuthbert et al. 2019a), detailed information on other aquifer properties is not consistently available across datasets, limiting a more in-depth assessment. Instead, the differing model performances between NA-N and NA-S, despite similar data quantities, may be partially explained by large-scale map products: wells in NA-N are often located in smaller, localized aquifers that may respond rapidly to rainfall, whereas those in NA-S are more often associated with larger regional groundwater systems (Richts et al. 2011). Aquifer responsiveness may significantly influence how effectively LSTM models can capture groundwater dynamics in response to changes of precipitation and temperature. Additionally, human activities – such as extensive unsustainable groundwater pumping, as observed in California (de Graaf et al. 2019;

Liu et al. 2022b) – can decouple GWLs from climate drivers, thereby reducing model performance. Similar effects have been observed in single-well regional modeling in Northern Germany, where lower performance was attributed to anthropogenic influences like proximity to waterworks (Gomez et al. 2024).

Lastly, the accuracy of the dynamic input data must be considered. In this study, inputs were solely derived from ERA5 reanalysis, which offers global consistency but may not match the resolution or precision of regionally available climate data. In some regions or at certain well sites, finer-resolution or more localized inputs might yield locally improved predictions.

5.5.5 Insights into spatial representation through embeddings

The embedding layer introduced in our LSTM model (see Section 5.4.3) enables exploration of how GWL dynamics are internally represented. By transforming well identifiers into trainable vectors, the model learns to differentiate between wells in a way that supports GWL prediction from climate inputs. This differentiation is learned directly from the time series data, without relying on externally provided static features. In the following, we examine whether the resulting embedding space may capture meaningful variation across wells that relates to underlying physical conditions.

Despite the inherently abstract nature of embeddings, we found discernible patterns in the reduced embedding space using PCA and t-SNE. Wells grouped by continent (Figure 18a) and by country within Europe (Figure 18b) show considerable overlap in the embedding space, supporting the findings from Section 5.5.3, where we hypothesized that the global model does not outperform regional models because a broad variety of GWL dynamics is already represented within regional datasets. Notably, wells from Europe (EU; blue dots), which dominate the dataset, display the widest dispersion in the embedding space, reflecting the high diversity in GWL dynamics across the continent. In contrast, wells from India (IN; black dots) cluster more tightly, suggesting more homogeneous behavior among the 200 Indian wells compared with the 1,000 wells from different parts of Europe. Among the visible groupings in the 2D t-SNE space, distinct subclusters appear, including those in the right corner containing wells from the Netherlands (blue dots in Figure 18a/b) and Australia (turquoise dots in Figure 18a). The partial alignment of embeddings with geographic groupings is expected to some extent given similarities in climate-driven GWL responses, as also suggested by the performance of climate-only models in Section 5.5.2. At the same time, the observed differences in overlap and separation between and within regions suggest that the embeddings may also capture site-specific functional variation,

such as differences in hydrograph behavior.

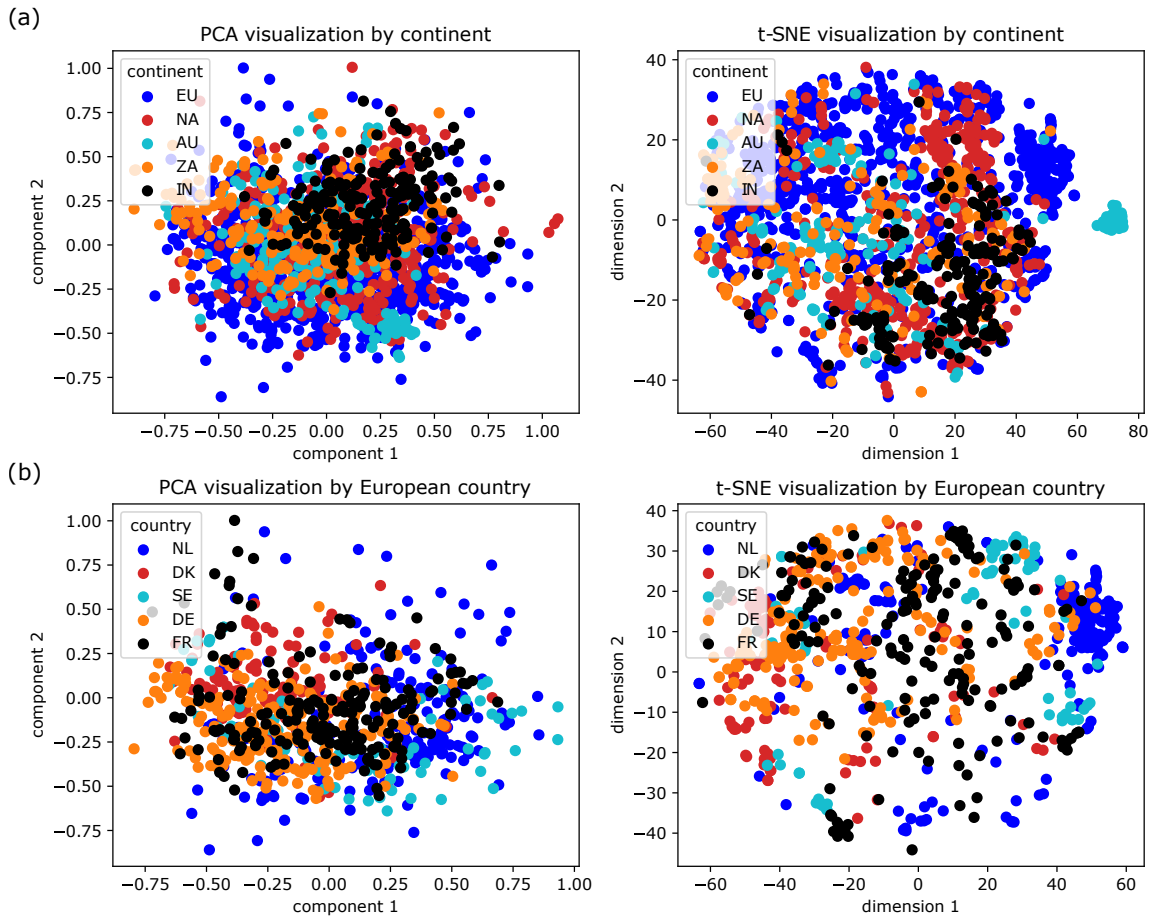


Figure 18: Visualization of the embedding space from the global model using PCA and t-SNE, colored by (a) continent – Europe (EU), North America (NA), Australia (AU), South Africa (ZA), and India (IN) – and (b) European country – Netherlands (NL), Denmark (DK), Sweden (SE), Germany (DE), and France (FR). Each point represents a well, and its position reflects relationships in GWL dynamics as learned by the model through embeddings.

To test this, we compared embedding-based clustering to hydrograph-based GWL clusters (see Section 5.4.4), which represent distinct temporal response patterns such as differences in seasonal amplitude, timing, and regularity (Nolte et al. 2024). The alignment between both clusterings was moderate, with silhouette scores between 0.17 and 0.21 and NMI scores between 0.24 and 0.31 across runs. While the relatively low silhouette scores indicate weak to moderate separation (Kaufman and Rousseeuw 2009), this is consistent with the original hydrograph-based clusters, which were not sharply delineated. The NMI scores offer further context: they are comparable to values reported in other hydrological studies, such as Gupta and Karthikeyan (2024), where similarly modest NMI scores

emerged when comparing clusters based on different drought-related features. In that context, the low agreement reflected both the fuzzy nature of functional types and the fact that different input features emphasize different aspects of system behavior. Likewise, in our case, embedding-based and hydrograph-based clusterings offer distinct but partially overlapping perspectives on groundwater dynamics.

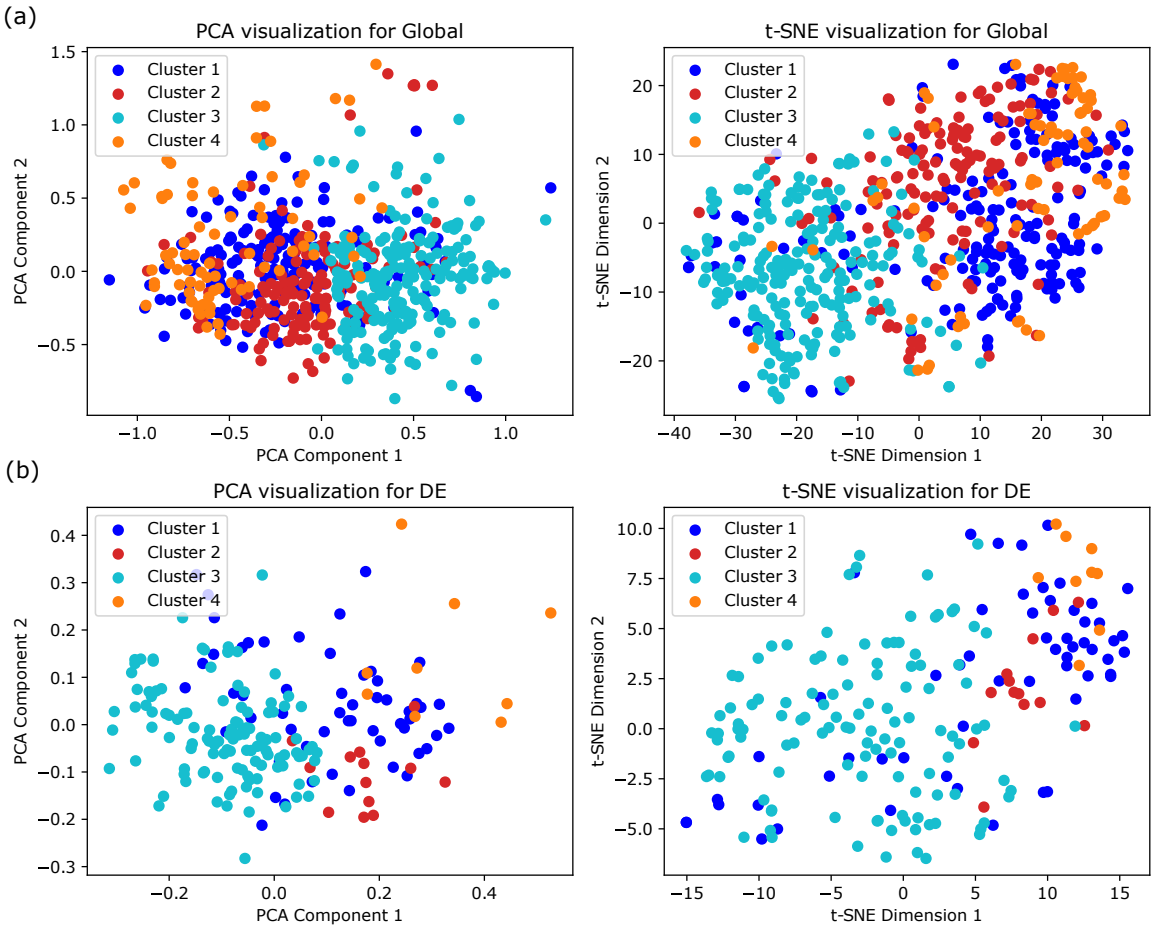


Figure 19: Visualization of the embedding space from (a) the global model and (b) the DE model using PCA and t-SNE. Wells that are shared between this study and Nolte et al. (2024) are plotted and color-coded according to four clusters representing distinct GWL dynamics, as previously defined using hydrograph analysis (Nolte et al. 2024). Each point represents a well, and its position reflects relationships in GWL dynamics as learned by the model through embeddings.

The partial alignment with hydrograph types, together with the visual patterns in Figure 19, suggests that the embedding space captures meaningful variation across wells related to underlying physical conditions such as climate, water table depth, and lithology (Nolte et al. 2024). In particular, the hydrograph-based clusters can be visually distinguished in the embedding space, forming partially separated regions that support

the idea that embeddings reflect functional similarity across wells. This occurred despite the model not receiving any explicit physical descriptors, highlighting the potential of embeddings to differentiate groundwater behavior beyond simple geographic correlation and to infer hydrogeological similarity directly from time series alone (Barthel et al. 2021).

While this suggests encouraging potential, the low silhouette and NMI values also underline a fundamental limitation: statistical clustering may struggle to capture functional (dis)similarity in groundwater systems, where dynamics often form gradual continua rather than discrete classes. Confirming hydrogeological similarity in embedding space therefore remains challenging without corresponding high-resolution or even site-specific environmental data. Future work should therefore aim to validate embedding interpretations using independent descriptors where available. Yet, recent studies highlight the limitations of this approach in practice: Chidepudi et al. (2025), working with wells across France, found that static environmental features often lack sufficient resolution or reliability to support meaningful interpretation using tools like SHAP, while Kunz et al. (2025), in a Germany-based study, observed only marginal performance gains from including static descriptors and noted that many such features fail to capture site-specific hydrogeological conditions, thus constraining both predictive value and interpretability.

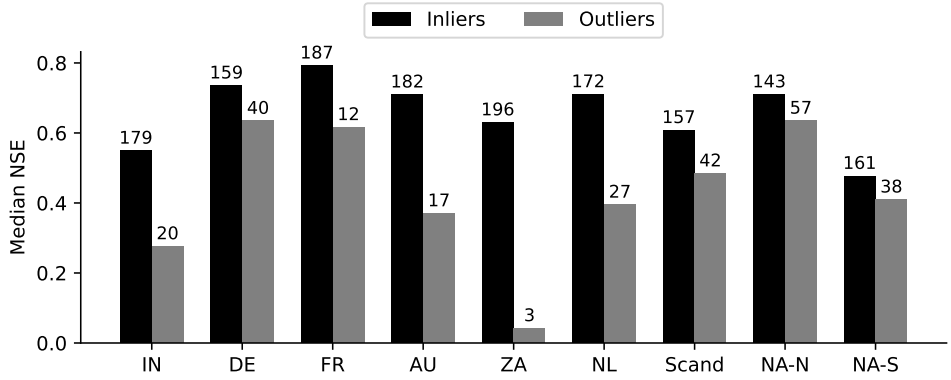


Figure 20: Median Nash-Sutcliffe Efficiency (NSE) scores for inlier and outlier wells across nine focus regions, based on DBSCAN clustering (with eps=0.25, min_samples=2) in the embedding space. Numbers above bars indicate the number of wells per group (inliers in black, outliers in gray). The comparison reveals a systematically lower model performance for wells identified as outliers in the embedding space.

The interpretability of embeddings depends not only on actual groundwater behavior but also on data quality. Sparse, short, irregular, or noisy time series can distort embedding formation, not due to hydrogeological distinctiveness, but because of reduced learnability.

This introduces a circular dependency: poor data quality can degrade embeddings, which in turn limits interpretability and obscures the underlying cause. To investigate this, we applied DBSCAN (see Section 5.4.4) to flag outliers in the embedding space and compared their predictive performance. Across regional models, these outlier wells consistently showed lower NSE scores than inliers (Figure 20), suggesting that they are either underrepresented in the training data or exhibit atypical behavior due to natural variability, anthropogenic influence, or poor data quality.

Looking ahead in terms of embedding implementation, evaluating embedding stability across model runs could provide deeper insight into the robustness of learned representations. Similarly, testing lower-dimensional embeddings, such as two or three dimensions, may offer more interpretable representations without loss of predictive performance. Preliminary tests (not shown) indicated that reducing from seven to three dimensions preserved nearly identical predictive skill, suggesting that compact representations (Bassi et al. 2024) may be sufficient for both modeling and interpretation.

5.6 Conclusions

This study advances large-scale GWL modeling in three key ways. First, it presents the first global, systematic evaluation of multi-site LSTM models across more than 1,800 globally distributed wells, using GWL time series with realistic variability in completeness, sampling frequency, and potential noise. Second, it demonstrates that trainable site identification (also called embeddings) can partially and effectively substitute for static environmental descriptors, which are often unavailable or uncertain, and which recent studies have shown to function more as unique site identifiers than physically meaningful predictors. Third, it provides the first analysis of the learned embedding space in a groundwater context, revealing emergent spatial and functional patterns that suggest the model captures some relevant hydrogeological structure even without explicit physical input. More broadly, the methods explored here may also support applications across the wider field of hydrology.

Embeddings provide a scalable and efficient alternative to one-hot encodings. Embedding-based models achieved high performance in data-rich regions, with median NSE scores exceeding 0.7 in countries like France and Germany. This indicates that the models not only work technically but also achieve a level of predictive accuracy that allows for practical application in regional analysis where data availability is sufficient. The analysis of the embedding space suggests that DL models can internally capture site-specific behavior

directly from time series data, which, in the case of GWL time series, can be related to differences in seasonal recharge dynamics, response timing, or longer-term variability – all of which may reflect underlying factors such as lithology or water table depth. The embeddings approach, therefore, holds promise not only for improving predictive performance but also for uncovering previously unrecognized patterns in groundwater systems, opening new ways for data-driven discovery. However, confidence in explaining these representations remains limited because this ultimately depends on the availability of independent physical descriptors, which may be explored through targeted case studies where such data exist.

The study also shows that global models perform comparably to regional models, even when integrating lower-quality data. This eliminates the need for separate region-specific or single-site models and highlights the potential for truly geographically global DL applications. However, a limitation is that adding high-quality GWL time series from other regions did not improve predictions in data-sparse regions. Achieving reliable performance in such settings remains a key challenge for multi-site DL models and warrants further investigation. These findings underscore the importance of acquiring sufficient data coverage, particularly at high temporal resolution, within the same regions for which predictive accuracy is currently limited.

Looking ahead, entity-aware multi-site models offer a promising direction for scalable time series prediction at continental to global scales. To support operational use cases such as real-time monitoring and climate impact assessments, future work should evaluate the model performance in true forecasting settings and further assess the stability and interpretability of site embeddings. Particular attention is needed for data-scarce or irregular time series, where strategies like transfer learning may help overcome current limitations in model accuracy. While embeddings cannot generalize to entirely unseen wells, they provide a powerful solution when static input features are missing or unreliable, and may be extended to new wells through targeted fine-tuning. Future work may also explore more advanced data fusion strategies beyond simple concatenation to better integrate static and dynamic information. Together, these directions will help advance DL toward becoming a practical and reliable tool for large-scale groundwater prediction under real-world data constraints.

6 Synthesis

Groundwater resources are central to water security and ecosystems, yet they are increasingly reshaped by groundwater abstraction, land-use change, and shifts in recharge under climate change. In coastal regions, these pressures act alongside changing boundary conditions at the land–sea interface. Until recently, little globally comparable, observation-based evidence existed on coastal groundwater systems. Within this thesis, an unprecedented compilation of about half a million coastal GWL monitoring locations worldwide was assembled and harmonized into the CGWL dataset (Chapter 2). The dataset was then used with the aim of advancing understanding of coastal groundwater dynamics and improving their prediction at global scale (Chapters 3 – 5). The main findings for the research questions (RQs) outlined in Section 1.4 are synthesized in the following sections. The chapter closes with overarching perspectives on future research.

6.1 Land–sea exchange and susceptibility to seawater intrusion

Chapter 3 addressed RQ1 (*How is land–sea exchange and susceptibility to seawater intrusion characterized along global coasts?*) by adding globally comparable, observation-based evidence to a state of knowledge that is dominated by case studies and top-down approaches that have uncertainties due to strong assumptions and sparse observational evaluation. Based on large stretches of monitored coastline, mainly across Europe, North and Central America, South Africa, India, and Oceania, the chapter shows where SWI susceptibility conditions occur, how far inland they extend, and how they evolve over decadal time scales. Across these monitored coastal areas, roughly one-third show land–sea hydraulic gradients indicative of reduced freshwater discharge potential and higher SWI susceptibility, and these conditions often extend beyond the immediate shoreline. Susceptibility patterns derived from hydraulic gradients combined with aridity as a recharge-limitation proxy broadly align with many previously documented SWI-affected regions from case studies (Cao et al. 2021; Kretschmer et al. 2025a) and with the only prior observation-based large-scale assessment across the United States (Jasechko et al. 2020). Beyond this consistency, a key advance is that coastlines with similar gradient-aridity configurations are captured more comprehensively at the global scale, while retaining strong within-region heterogeneity that is often masked in spatially smoothed large-scale assessments. In several regions, susceptibility conditions form contiguous coastal belts, yet patterns can vary sharply over short distances. The results also indicate that susceptibility is dynamic where coastal head conditions are changing.

Across 1990-2024, 21% of monitored coastal areas show significant GWL changes of at least ± 0.1 m yr^{-1}, with a larger share in arid regions and deeper groundwater systems where longer response times can delay recovery once heads decline and can prolong periods in which SWI-relevant conditions persist. The SWI case-study literature is skewed toward urban, economically developed coasts, but Chapter 3 suggests that head changes do not follow population density globally and that declines often occur in rural coastal areas. This mismatch motivates stronger attention to rural coasts where both agricultural abstraction and off-site pumping for urban supply are likely to drive head declines. Importantly, in the immediate shoreline zone, susceptibility and significant head change often co-occur. Assuming that recent monotonic GWL trends continue, most sites already in susceptible conditions appear likely to remain so over the next decade (93.4%), while GWL declines could introduce susceptible conditions at 3.5% of total sites evaluated. This is only slightly more frequent than the share of sites where rising GWLs would move them out of susceptibility. However, continued sea-level rise and increasing drought and abstraction pressure are likely to shift additional coastlines toward SWI-susceptible gradient states. What can be inferred from Chapter 3 remains strongly constrained by current monitoring data availability and contextual information. Head-based average conditions and monotonic trends do not resolve seasonal or event-driven SWI-prone states that require regular, high-resolution time series, and some records are too short to robustly distinguish persistent change from multi-annual variability. Translating susceptibility into current and future SWI vulnerability also requires co-located salinity monitoring and hydrogeologic context, together with information on nonlinear shifts or management interventions that cannot be inferred from head trends alone. For large-scale, observation-based groundwater assessments that often rely on simple trend calculations, Chapter 3 also demonstrates that reporting trends across multiple time windows can serve as a practical robustness check of trend direction and magnitude when long records are unavailable. Across windows, trend direction agreed for 76%, while clearly inconsistent opposite-sign trends with substantial magnitudes in both windows occurred in about 5% and are concentrated in a small number of regions. Especially in these regions, regional trend narratives can depend on the chosen period and window length. Nevertheless, the results support anticipating risks to water security and ecosystems in coastal regions and help prioritize monitoring and management through the quantitative screening provided. They also strengthen process understanding of land–sea exchange by providing observation-based constraints on coastal head and gradient conditions that can also be used to evaluate, interpret, and improve large-scale modeling.

6.2 Groundwater dynamic regimes and explanations

Chapter 4 addressed RQ2 (*What types of recurring regimes of coastal groundwater dynamics emerge globally, and to what extent can they be explained?*) by closing a key gap: although local and regional studies document variability in GWL dynamics, it has remained unclear whether coastal GWL hydrographs exhibit recurring, comparable patterns globally. Building on the idea that hydrograph classification can reveal hydrogeological similarity (Heudorfer et al. 2019; Barthel et al. 2021), Chapter 4 provided the first systematic classification of coastal groundwater dynamics at the global scale and quantified the extent to which coastal groundwater behavior includes recurring dynamic regimes, rather than being entirely site-specific. When summarizing thousands of coastal GWL hydrographs, four recurring archetypal dynamic regimes emerge across continents and climate zones. The resulting typology offers a scalable reference for global comparison and for assessing hydrogeological similarity across scales. Wells in close proximity can show different dynamics, while similar regimes recur across distant regions, highlighting the potential to transfer expectations about GWL behavior beyond local geographic context. Groundwater depth and topography provide a useful first-order partitioning by distinguishing a damped, low-responsiveness regime from the other three, but because shallow aquifers span multiple regimes, depth cannot be used for transferring expectations about full dynamics. Beyond this coarse split, climate and soil/lithology help differentiate regimes. Within the available descriptor set and explainable ML model, the strongest differentiation comes from variables closer to the governing processes (e.g., temperature compared to derived indices such as the aridity index), suggesting that, given the currently available global descriptor set, data-driven GWL modeling and prediction benefit most from physically direct, higher-information predictors that reflect the main controls, while many additional indirect or coarsely resolved descriptors add limited information. Together with the moderate ability to predict regime membership from the available descriptors (62% accuracy across four regimes), the results indicate that important explanatory information is still missing at the global scale. Especially groundwater-relevant subsurface information (e.g., aquifer heterogeneity/architecture and connectivity) and time-varying or local drivers, including localized groundwater use, is difficult to represent and disentangle with current global proxies. The results imply that currently available global descriptor datasets are not yet sufficient to enable broadly transferable large-scale modeling of GWL dynamics, which constrains regionalization to unmonitored regions and limits generalizable prediction. At the same time, regime classification is not a trivial prediction task. Regime signatures overlap, so some wells fall near regime boundaries and cannot be assigned unambiguously, even with perfect

information. This reflects the high-dimensional complexity of groundwater dynamics and highlights the need for methods that can accommodate it in prediction settings. Which regime information is most useful depends on the application. For example, the outcomes from Chapter 4 can help stratify monitoring networks at the global scale by general response type. Specifically in the context of RQ1, regime assignments could complement mean and trend diagnostics of SWI susceptibility conditions by adding dynamic context for the subset of wells with sufficiently dense time series. For large-scale modeling and evaluation, regimes provide pattern-based targets where models can be assessed on whether they reproduce characteristic dynamic behaviors rather than only minimizing overall magnitude-based error scores.

6.3 Predictive modeling of groundwater levels at the global scale

Chapter 5 addressed RQ3 (*How can coastal groundwater levels be predicted at the global scale with deep learning under realistic data limitations?*) by benchmarking a multi-site state-of-the-art DL model across 1,800 coastal wells from nine regions worldwide for weekly-ahead GWL prediction. The novelty lies in moving beyond typical single-region, site-specific DL applications by explicitly testing scalability at the global scale, with the aim of improving the efficiency and usability of data-driven prediction for both research and practice. Multi-site DL typically relies on static site descriptors to encode inter-site (dis)similarity (Kratzert et al. 2019), yet in hydrogeology such descriptors are often missing, inconsistent, or only weak proxies, also supported by the results of Chapter 4. Chapter 5 showed that high prediction skill can be achieved in regions with high GWL data availability (i.e., relatively dense observations with fewer gaps) using only dynamic climatic inputs together with site identifiers. Moreover, a jointly trained global model matched the skill of region-specific models, indicating that pooled multi-region training can serve as a scalable default without compromising regional performance. Both results lower the effort to develop and maintain models across large monitoring networks. They support global-scale, bottom-up data-driven modeling as a practical complement to existing groundwater modeling and assessment approaches. However, adding training data from data-rich regions did not consistently improve prediction in data-scarce regions. Hence, within the tested multi-site setup using the widely used LSTM model class, cross-regional generalization remains limited and improving predictions under monitoring data scarcity cannot be achieved simply by pooling more data elsewhere. Predictive skill varied strongly across regions and was primarily constrained by the

quantity and, in particular, the sampling frequency and regularity of available GWL observations. Performance was high in data-rich networks (e.g., France and Germany) but dropped markedly under sparse or irregular GWL time series. This indicates that, under current monitoring constraints, operationally reliable prediction at the global scale is feasible only for a subset of monitoring networks, and emphasizes the value of regular sampling for monitoring designs that aim to support prediction. Methodological advances such as transfer learning or targeted fine-tuning may help reduce data requirements by leveraging shared climate–response patterns, but cannot fully compensate for very sparse or irregular GWL observations in the absence of physically meaningful site descriptors (covariates). Another limitation hindering the practical use of DL models in geosciences is that predictive accuracy alone does not guarantee physically meaningful learning. Models therefore need to be interpretable, including through explainable Machine Learning (XAI) approaches, so that users can assess whether predictions are based on plausible relationships, remain robust under non-stationarity, and are hence suitable for scientific inference and decision support. In Chapter 5, learned site representations (embeddings) provided a compact way to capture site specificity and enabled post hoc checks of cross-site similarity and model behavior. The learned embedding space exhibited emergent geographical and functional structure, with partial alignment to the hydrograph-based regimes derived in Chapter 4. While such representations are promising, their meaning is difficult to validate independently under current data constraints. Thus, achieving useful skill without static descriptors does not remove the need for physically grounded information to derive explanations of modeling results. This motivates targeted case studies in data- and context-rich regions to further test whether the embedding space recovers physically meaningful groupings of wells (i.e., similarity in hydrogeologic setting or dominant forcing). Beyond the hindcast evaluation in this thesis, it remains to be tested whether multi-site DL models remain reliable in true forecasting settings, including under non-stationary conditions such as changing climate and pumping, and whether multi-site training improves generalization of climate-related impacts (e.g., drought responses) by transferring learned relationships from previously observed conditions to regions or periods where such conditions newly emerge.

6.4 Conclusions and outlook

Across all three research questions, this thesis demonstrated that harmonized, large-sample coastal GWL observations can turn fragmented monitoring into globally comparable, locally faithful evidence on coastal groundwater state, dynamics, and predictability. The resulting standardized diagnostics and benchmarks position bottom-up, observation-driven

assessment as a practical complement to spatially continuous but assumption-dependent top-down estimates and to process-resolving yet less comparable case studies. At the same time, the global-scale focus precludes causal attribution, and the results necessarily reflect the monitoring footprint, data availability, and metadata limitations of available records.

Further progress toward a more globally consistent understanding of coastal groundwater systems and toward reliable prediction at scale hinges first on strengthening the data foundation. Priorities in sustained and expanded coastal monitoring include preserving relevant dynamics through sufficiently regular sampling, consistent measurement conventions, and standardized, openly accessible time series, metadata, and hydrogeologic context that constrains dynamics. Monitoring expansion should particularly target coastline segments that remain sparsely represented or missing in the CGWL dataset, including most of Africa, equatorial regions, large parts of South America and Asia, and many small islands. Integrating co-located water-quality observations is essential for comprehensive freshwater-availability assessment, because, for example, without salinity indicators head- and gradient-based diagnostics remain proxies that cannot confirm when or where susceptibility translates into salinization impacts. Improved documentation of human pressures (e.g., abstraction) is equally critical to separate anthropogenic impacts from climatic drivers and to interpret head trends, dynamic regimes, and model predictions and forecasts. In general, globally consistent descriptor datasets need to be further developed to provide more meaningful, groundwater-relevant information that better supports explanation, transferability, and model evaluation at scale.

Because such comprehensive information will remain unevenly available globally in the near term, a pragmatic way forward for the bottom-up pathway adopted in this thesis is to establish a distributed set of sentinel, comprehensively characterized coastal monitoring sites for long-term tracking and coordinated reporting. Guided by the diagnostics developed in this thesis (e.g., spanning dynamic regimes and including SWI-prone settings), these sites can serve as representative reference points for broader groups of similar coastal settings. Methodologically, this thesis points to several directions for advancing coastal groundwater assessment and predictive modeling at scale. One priority is the tighter coupling of bottom-up diagnostics with top-down modeling. Spatially complete approaches should be evaluated against large samples of newly available data on observed head conditions, land–sea gradients, and characteristic dynamic behaviors. A second direction is that scalable DL prediction should prioritize focusing on approaches that remain robust under sparse and irregular monitoring and that are interpretable, so that skill can be assessed alongside plausibility and reliability across heterogeneous

settings and under non-stationarity. This is essential if DL models are to become decision-relevant tools for practitioners and credible, hypothesis-generating approaches for process understanding worldwide. More broadly, progress will likely benefit from closer integration of process-based structure, observation-based diagnostics, and data-driven learning. For example, different modeling paradigms are increasingly combined in hybrid frameworks that leverage their complementary strengths. Ultimately, observation-constrained large-scale models could improve the representation of groundwater interactions with surface water, wetlands, and the coastal ocean and help disentangle the effects of abstraction and land use from climate-driven shifts in recharge and sea level, enabling more decision-relevant assessments of future groundwater availability and vulnerability at the global scale.

Bibliography

Abhik, S., D. Dommenget, S. McGregor, D. K. Hutchinson, S. Steinig, J. Zhu, F. A. Capitanio, D. J. Lunt, I. Niezgodzki, and G. Knorr (2025). "Stronger and prolonged El Niño-Southern Oscillation in the Early Eocene warmth". In: *Nature Communications* 16.1, 4053. 10.1038/s41467-025-59263-7.

Adams, K. H., J. T. Reager, B. A. Buzzanga, C. H. David, A. H. Sawyer, and B. D. Hamlington (2024). "Climate-induced saltwater intrusion in 2100: Recharge-driven severity, sea level-driven prevalence". In: *Geophysical Research Letters* 51.22, e2024GL110359. 10.1029/2024GL110359.

Addor, N., H. X. Do, C. Alvarez-Garreton, G. Coxon, K. Fowler, and P. A. Mendoza (2020). "Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges". In: *Hydrological Sciences Journal* 65.5, 712–725. 10.1080/02626667.2019.1683182.

Addor, N., A. J. Newman, N. Mizukami, and M. P. Clark (2017). "The CAMELS data set: catchment attributes and meteorology for large-sample studies". In: *Hydrology and Earth System Sciences* 21.10, 5293–5313. 10.5194/hess-21-5293-2017.

Aeschbach-Hertig, W. and T. Gleeson (2012). "Regional strategies for the accelerating global problem of groundwater depletion". In: *Nature Geoscience* 5.12, 853–861. 10.1038/ngeo1617.

Akbik, A., D. Blythe, and R. Vollgraf (2018). "Contextual string embeddings for sequence labeling". In: *Proceedings of the 27th international conference on computational linguistics*, 1638–1649.

Alfarrah, N. and K. Walraevens (2018). "Groundwater overexploitation and seawater intrusion in coastal areas of arid and semi-arid regions". In: *Water* 10.2, 143. 10.3390/w10020143.

Alley, W. M., R. W. Healy, J. W. LaBaugh, and T. E. Reilly (2002). "Flow and storage in groundwater systems". In: *Science* 296.5575, 1985–1990. 10.1126/science.1067123.

Alley, W. M., T. E. Reilly, and O. L. Franke (1999). Sustainability of ground-water resources. Vol. 1186. U.S. Geological Survey Circular. Denver, Colorado: U.S. Geological Survey.

AQUASTAT (2021). Percentage of irrigated area serviced by groundwater (Global): [data set]. URL: https://data.apps.fao.org/catalog/dataset/d49db282-7c50-46c2-b210-3e197d767da3 (visited on 01/05/2024).

Arifin, A., M. Shamsudduha, Ramdhan, A. Rateb, B. R. Scanlon, T. Setiawan, M. I. Iman, A. Husna, and R. G. Taylor (2025). "Plausibility criteria for GRACE–derived groundwater storage changes from aquifers globally". In: *Geophysical Research Letters* 52.22, e2025GL118580. 10.1029/2025GL118580.

Ascott, M. J., D. M. J. Macdonald, E. Black, A. Verhoef, P. Nakohoun, J. Tirogo, W. J. P. Sandwidi, J. Bliefernicht, J. P. R. Sorensen, and A. Y. Bossa (2020). "In situ observations and lumped parameter model reconstructions reveal intra-annual to multidecadal variability in groundwater levels in sub-Saharan Africa". In: *Water Resources Research* 56.12, e2020WR028056. 10.1029/2020WR028056.

Babre, A., A. Kalvāns, Z. Avotniece, I. Retiķe, J. Bikše, K. Popovs, M. Jemeljanova, A. Zelenkevičs, and A. Dēliņa (2022). "The use of predefined drought indices for the assessment of groundwater drought episodes in the Baltic States over the period 1989–2018". In: *Journal of Hydrology: Regional Studies* 40, 101049. 10.1016/j.ejrh.2022.101049.

Balacumaresan, H., M. A. Imteaz, M. A. Aziz, and T. Choudhury (2024). "Use of artificial intelligence modelling for the dynamic simulation of urban catchment runoff". In: *Water Resources Management* 38.10, 3657–3683. 10.1007/s11269-024-03833-9.

Barbarossa, V., M. A. J. Huijbregts, A. H. W. Beusen, H. E. Beck, H. King, and A. M. Schipper (2018). FLO1K, global maps of mean, maximum and minimum annual streamflow at 1 km resolution from 1960 through 2015: [data set]. 10.6084/m9.figshare.c.3890224.v1.

Barlow, P. M. and E. G. Reichard (2010). "Saltwater intrusion in coastal regions of North America". In: *Hydrogeology Journal* 18.1, 247–260. 10.1007/s10040-009-0514-3.

Barthel, R., E. Haaf, M. Giese, M. Nygren, B. Heudorfer, and K. Stahl (2021). "Similarity-based approaches in hydrogeology: proposal of a new concept for data-scarce groundwater resource characterization and prediction". In: *Hydrogeology Journal* 24.4, 3392. 10.1007/s10040-021-02358-4.

Barthel, R., E. Haaf, M. Nygren, and M. Giese (2022). "Systematic visual analysis of groundwater hydrographs: potential benefits and challenges". In: *Hydrogeology Journal* 30.2, 359–378. 10.1007/s10040-021-02433-w.

Bassi, A., M. Höge, A. Mira, F. Fenicia, and C. Albert (2024). "Learning landscape features from streamflow with autoencoders". In: *Hydrology and Earth System Sciences* 28, 4971–4988. 10.5194/hess-28-4971-2024.

Baulon, L., D. Allier, N. Massei, H. Bessiere, M. Fournier, and V. Bault (2022a). "Influence of low-frequency variability on groundwater level trends". In: *Journal of Hydrology* 606, 127436. 10.1016/j.jhydrol.2022.127436.

Baulon, L., N. Massei, D. Allier, M. Fournier, and H. Bessiere (2022b). "Influence of low-frequency variability on high and low groundwater levels: example of aquifers in the Paris Basin". In: *Hydrology and Earth System Sciences* 26.11, 2829–2854. 10.5194/hess-26-2829-2022.

Bear, J. (1979). Hydraulics of groundwater. New York: McGraw-Hill.

Beck, H. E., A. I. J. M. van Dijk, A. de Roo, D. G. Miralles, T. R. McVicar, J. Schellekens, and L. A. Bruijnzeel (2016). "Global-scale regionalization of hydrologic model parameters". In: *Water Resources Research* 52.5, 3599–3622. 10.1002/2015WR018247.

Bender, S. (2007). "Die Aussageunschärfe bei der Verwendung heterogener Datensätze im Rahmen wasserwirtschaftlicher Fragestellungen". In: *Bochumer Geowissenschaftliche Arbeiten* 8, 99.

Berendrecht, W., M. van Vliet, and J. Griffioen (2022). "Combining statistical methods for detecting potential outliers in groundwater quality time series". In: *Environmental monitoring and assessment* 195.1, 85. 10.1007/s10661-022-10661-0.

Bergen, K. J., P. A. Johnson, M. V. de Hoop, and G. C. Beroza (2019). "Machine learning for data-driven discovery in solid Earth geoscience". In: *Science* 363.6433, eaau0323. 10.1126/science.aau0323.

Berghuijs, W. R., R. A. Collenteur, S. Jasechko, F. Jaramillo, E. Luijendijk, C. Moeck, Y. van der Velde, and S. T. Allen (2024). "Groundwater recharge is sensitive to changing long-term aridity". In: *Nature Climate Change* 14.4, 357–363. 10.1038/s41558-024-01953-z.

Bloomfield, J. P. and B. P. Marchant (2013). "Analysis of groundwater drought building on the standardised precipitation index approach". In: *Hydrology and Earth System Sciences* 17.12, 4769–4787. 10.5194/hess-17-4769-2013.

Bloomfield, J. P., B. P. Marchant, and A. A. McKenzie (2019). "Changes in groundwater drought associated with anthropogenic warming". In: *Hydrology and Earth System Sciences* 23.3, 1393–1408. 10.5194/hess-23-1393-2019.

Blöschl, G. and M. Sivapalan (1995). "Scale issues in hydrological modelling: a review". In: *Hydrological Processes* 9.3–4, 251–290. 10.1002/hyp.3360090305.

Blöschl, G. et al. (2019). "Twenty-three unsolved problems in hydrology (UPH) – a community perspective". In: *Hydrological Sciences Journal* 64.10, 1141–1158. 10.1080/02626667.2019.1620507.

Blumstock, M., D. Tetzlaff, J. J. Dick, G. Nuetzmann, and C. Soulsby (2016). "Spatial organization of groundwater dynamics and streamflow response from different hydropedological units in a montane catchment". In: *Hydrological Processes* 30.21, 3735–3753. 10.1002/hyp.10848.

Börker, J., J. Hartmann, T. Amann, and G. Romero-Mujalli (2018). "Terrestrial sediments of the earth: development of a global unconsolidated sediments map database (GUM)". In: *Geochemistry, Geophysics, Geosystems* 19.4, 997–1024. 10.1002/2017GC007273.

Bosserelle, A. L., L. K. Morgan, D. E. Dempsey, and I. Setiawan (2023). "Shallow groundwater characterisation and hydrograph classification in the coastal city of Ōtautahi/Christchurch, New Zealand". In: *Hydrogeology Journal* 32, 577–600. 10.1007/s10040-023-02745-z.

Bosserelle, A. L., L. K. Morgan, and M. W. Hughes (2022). "Groundwater rise and associated flooding in coastal settlements due to sea–level rise: a review of processes and methods". In: *Earth's Future* 10.7, e2021EF002580. 10.1029/2021EF002580.

Botterill, T. E. and H. K. McMillan (2023). "Using machine learning to identify hydrologic signatures with an encoder–decoder framework". In: *Water Resources Research* 59.3. 10.1029/2022WR033091.

Breiman, L. (2001). "Random Forests". In: *Machine learning* 45, 5–32. 10.1023/A:1010933404324.

Caliński, T. and J. Harabasz (1974). "A dendrite method for cluster analysis". In: *Communications in Statistics* 3.1, 1–27. 10.1080/03610927408827101.

Campello, R. J., P. Kröger, J. Sander, and A. Zimek (2020). "Density-based clustering". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.2, e1343. 10.1002/widm.1343.

Cao, T., D. Han, and X. Song (2021). "Past, present, and future of global seawater intrusion research: A bibliometric analysis". In: *Journal of Hydrology* 603, 126844. 10.1016/j.jhydrol.2021.126844.

Chávez García Silva, R., R. Reinecke, N. K. Copty, D. A. Barry, E. Heggy, D. Labat, P. P. Roggero, D. Borchardt, M. Rode, and J. J. Gómez-Hernández (2024). "Multi-decadal groundwater observations reveal surprisingly stable levels in southwestern Europe". In: *Communications Earth & Environment* 5, 387. 10.1038/s43247-024-01554-w.

Chidepudi, S. K. R., N. Massei, A. Jardani, B. Dieppois, A. Henriot, and M. Fournier (2025). "Training deep learning models with a multi-station approach and static aquifer attributes for groundwater level simulation: what is the best way to leverage regionalised information?" In: *Hydrology and Earth System Sciences* 29.4, 841–861. 10.5194/hess-29-841-2025.

Chidepudi, S. K. R., N. Massei, A. Jardani, A. Henriot, D. Allier, and L. Baulon (2023). "A wavelet-assisted deep learning approach for simulating groundwater levels affected by low-frequency variability". In: *Science of the Total Environment* 865, 161035. 10.1016/j.scitotenv.2022.161035.

CIESIN (2018). Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11 (Version 4.11): [data set]. 10.7927/H49C6VHW.

Clark, S. R., D. Pagendam, and L. Ryan (2022). "Forecasting multiple groundwater time series with local and global deep learning networks". In: *International Journal of Environmental Research and Public Health* 19.9, 5091. 10.3390/ijerph19095091.

Condon, L. E., S. Kollet, M. F. P. Bierkens, G. E. Fogg, R. M. Maxwell, M. C. Hill, H.-.-J. H. Fransen, A. Verhoef, A. F. van Loon, M. Sulis, and C. Abesser (2021). "Global groundwater modeling and monitoring: Opportunities and challenges". In: *Water Resources Research* 57.12, e2020WR029500. 10.1029/2020WR029500.

Cosby, A. G., V. Lebakula, C. N. Smith, D. W. Wanik, K. Bergene, Rose, D. Swanson, and Bloom (2024). "Accelerating growth of human coastal populations at the global and continent levels: 2000–2018". In: *Scientific Reports* 14, 22489. 10.1038/s41598-024-73287-x.

Costall, A. R., B. D. Harris, B. Teo, R. Schaa, F. M. Wagner, and J. P. Pigois (2020). "Groundwater throughflow and seawater intrusion in high quality coastal aquifers". In: *Scientific Reports* 10, 9866. 10.1038/s41598-020-66516-6.

Costantini, M., J. Colin, and B. Decharme (2023). "Projected climate–driven changes of water table depth in the world's major groundwater basins". In: *Earth's Future* 11.3, e2022EF003068. 10.1029/2022EF003068.

Cuthbert, M. O., T. Gleeson, N. Moosdorf, K. M. Befus, A. Schneider, J. Hartmann, and B. Lehner (2019a). "Global patterns and dynamics of climate–groundwater interactions". In: *Nature Climate Change* 9.2, 137–141. 10.1038/s41558-018-0386-4.

Cuthbert, M. O., R. G. Taylor, G. Favreau, M. C. Todd, M. Shamsudduha, K. G. Villholth, A. M. MacDonald, B. R. Scanlon, D. V. Kotchoni, and J.-M. Vouillamoz (2019b). "Observed

controls on resilience of groundwater to climate variability in sub-Saharan Africa". In: *Nature* 572.7768, 230–234. 10.1038/s41586-019-1441-7.

Dams, J., E. Salvadore, T. van Daele, V. Ntegeka, P. Willems, and O. Batelaan (2012). "Spatio-temporal impact of climate change on the groundwater system". In: *Hydrology and Earth System Sciences* 16.5, 1517–1531. 10.5194/hess-16-1517-2012.

Darcy, H. (1856). Les fontaines publiques de la ville de Dijon: Exposition et application des principes à suivre et des formules à employer dans les questions de distribution d'eau: Ouvrage terminé par un appendice relatif aux fournitures d'eau de plusieurs villes, au filtrage des eaux et à la fabrication des tuyaux de fonte, de plomb, de tôle et de bitume. Victor Dalmont.

Davies, D. L. and D. W. Bouldin (1979). "A cluster separation measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2, 224–227. 10.1109/TPAMI.1979.4766909.

de Graaf, I., T. Gleeson, L. P. H. van Rens Beek, E. H. Sutanudjaja, and M. F. P. Bierkens (2019). "Environmental flow limits to global groundwater pumping". In: *Nature* 574.7776, 90–94. 10.1038/s41586-019-1594-4.

Díaz-Alcaide, S. and P. Martínez-Santos (2019). "Review: Advances in groundwater potential mapping". In: *Hydrogeology Journal* 27.7, 2307–2324. 10.1007/s10040-019-02001-3.

Diersch, H.-J. G. (2013). FEFLOW: finite element modeling of flow, mass and heat transport in porous and fractured media. Springer Science & Business Media.

Donnelly, C., K. Ernst, and B. Arheimer (2018). "A comparison of hydrological climate services at different scales by users and scientists". In: *Climate Services* 11, 24–35. 10.1016/j.cliser.2018.06.002.

Dyring, M., H. Hofmann, D. Stanton, P. Moss, and R. Froend (2022). "Ecohydrology of coastal aquifers in humid environments and implications of a drying climate". In: *Ecohydrology* 16.2, e2491. 10.1002/eco.2491.

Ebeling, P., R. Kumar, S. R. Lutz, T. Nguyen, F. Sarrazin, M. Weber, O. Büttner, S. Attinger, and A. Musolff (2022). "Water quality, discharge and catchment attributes for large-sample studies in Germany – QUADICA". In: *Earth System Science Data* 14.8, 3715–3741. 10.5194/essd-2022-6.

Elshall, A. S., A. D. Arik, A. I. El-Kadi, S. Pierce, M. Ye, K. M. Burnett, C. A. Wada, L. L. Bremer, and G. Chun (2020). "Groundwater sustainability: a review of the interactions between science and policy". In: *Environmental Research Letters* 15.9, 093004. 10.1088/1748-9326/ab8e8c.

Esri (2022). ArcGIS Pro, version 3.0.3: [software]. URL: https://www.esri.com.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 34, 226–231.

Eurostat (2021). Applying the Degree of Urbanisation: A methodological manual to define cities, towns and rural areas for international comparisons. Luxembourg: Publications Office of the European Union. 10.2785/706535.

Famiglietti, J. S. (2014). "The global groundwater crisis". In: *Nature Climate Change* 4.11, 945–948. 10.1038/nclimate2425.

Famiglietti, J. S. and M. Rodell (2013). "Water in the balance". In: *Science* 340.6138, 1300–1301. 10.1126/science.1240843.

Fan, X., T. J. Peterson, B. J. Henley, and M. Arora (2023a). "Groundwater sensitivity to climate variations across Australia". In: *Water Resources Research* 59.11, e2023WR035036. 10.1029/2023WR035036.

Fan, Y., H. Li, and G. Miguez-Macho (2013). "Global patterns of groundwater table depth". In: *Science* 339.6122, 940–943. 10.1126/science.1229881.

Fan, Y., H. Li, and G. Miguez-Macho (2023b). Global patterns of groundwater table depth: [data set]. URL: http://thredds-gfnl.usc.es/thredds/catalog/GLOBALWTDFTP/catalog.html (visited on 03/08/2023).

Fang, K., D. Kifer, K. Lawson, D. Feng, and C. Shen (2022). "The data synergy effects of time–series deep learning models in hydrology". In: *Water Resources Research* 58.4, e2021WR029583. 10.1029/2021WR029583.

Farr, T. G., P. A. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, M. Kobrick, M. Paller, E. Rodriguez, and L. Roth (2007). "The shuttle radar topography mission". In: *Reviews of Geophysics* 45.2. 10.1029/2005RG000183.

Ferguson, G. and T. Gleeson (2012). "Vulnerability of coastal aquifers to groundwater use and climate change". In: *Nature Climate Change* 2.5, 342–345. 10.1038/nclimate1413.

Foster, S., J. Chilton, G.-J. Nijsten, and A. Richts (2013). "Groundwater—a global focus on the 'local resource'". In: *Current Opinion in Environmental Sustainability* 5.6, 685–695. 10.1016/j.cosust.2013.10.010.

Frame, J. M., F. Kratzert, D. Klotz, M. Gauch, G. Shalev, O. Gilon, L. M. Qualls, H. V. Gupta, and G. S. Nearing (2022). "Deep learning rainfall–runoff predictions of extreme events". In: *Hydrology and Earth System Sciences* 26.13, 3377–3392. 10.5194/hess-26-3377-2022.

Frappart, F. and G. Ramillien (2018). "Monitoring groundwater storage changes using the Gravity Recovery and Climate Experiment (GRACE) satellite mission: A review". In: *Remote Sensing* 10.6, 829. 10.3390/rs10060829.

GebreEgziabher, M., D. Perrone, and S. Jasechko (2025). Undrainable pore spaces comprise half of US groundwater storage: [preprint]. 10.31223/X50Q8P.

Ghirardelli, A., E. Straffelini, E. Park, V. D'Agostino, R. Masin, and P. Tarolli (2024). "Global impact of seawater intrusion on coastal agriculture". In: *Environmental Research Letters* 20, 013005. 10.1088/1748-9326/ad9bcd.

Giese, M., E. Haaf, B. Heudorfer, and R. Barthel (2020). "Comparative hydrogeology – reference analysis of groundwater dynamics from neighbouring observation wells". In: *Hydrological Sciences Journal* 65.10, 1685–1706. 10.1080/02626667.2020.1762888.

Gleeson, T., M. Cuthbert, G. Ferguson, and D. Perrone (2020a). "Global groundwater sustainability, resources, and systems in the Anthropocene". In: *Annual Review of Earth and Planetary Sciences* 48, 431–463. 10.1146/annurev-earth-071719-055251.

Gleeson, T., T. Wagener, P. Döll, S. C. Zipper, C. West, Y. Wada, R. Taylor, B. Scanlon, R. Rosolem, S. Rahman, N. Oshinlaja, R. Maxwell, M.-H. Lo, H. Kim, M. Hill, A. Hartmann, G. Fogg, J. S. Famiglietti, A. Ducharne, I. de Graaf, M. Cuthbert, L. Condon, E. Bresciani, and M. F. P. Bierkens (2021). "GMD perspective: The quest to improve the evaluation of groundwater representation in continental- to global-scale models". In: *Geoscientific Model Development* 14.12, 7545–7571. 10.5194/gmd-14-7545-2021.

Gleeson, T., L. Wang–Erlandsson, M. Porkka, S. C. Zipper, F. Jaramillo, D. Gerten, I. Fetzer, S. E. Cornell, L. Piemontese, and L. J. Gordon (2020b). "Illuminating water cycle modifications and Earth system resilience in the Anthropocene". In: *Water Resources Research* 56.4, e2019WR024957. 10.1029/2019WR024957.

Gomez, M., M. Nölscher, A. Hartmann, and S. Broda (2024). "Assessing groundwater level modelling using a 1-D convolutional neural network (CNN): linking model performances to geospatial and time series features". In: *Hydrology and Earth System Sciences* 28.19, 4407–4425. 10.5194/hess-28-4407-2024.

Güler, C., M. A. Kurt, M. Alpaslan, and C. Akbulut (2012). "Assessment of the impact of anthropogenic activities on the groundwater hydrology and chemistry in Tarsus coastal plain (Mersin, SE Turkey) using fuzzy clustering, multivariate statistics and GIS techniques". In: *Journal of Hydrology* 414, 435–451. 10.1016/j.jhydrol.2011.11.021.

Guppy, L., P. Uyttendaele, K. G. Villholth, and V. Smakhtin (2018). Groundwater and Sustainable Development Goals: Analysis of Interlinkages. Vol. 4. UNU-INWEH Report Series. United Nations University. 10.53328/JRLH1810.

Gupta, A. and L. Karthikeyan (2024). "Role of initial conditions and meteorological drought in soil moisture drought propagation: An event–based causal analysis over South Asia". In: *Earth's Future* 12.10, e2024EF004674. 10.1029/2024EF004674.

Gupta, H. V., C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian (2014). "Large-sample hydrology: a need to balance depth with breadth". In: *Hydrology and Earth System Sciences* 18.2, 463–477. 10.5194/hess-18-463-2014.

Haaf, E., M. Giese, T. Reimann, and R. Barthel (2023). "Data-driven estimation of groundwater level time-series at unmonitored sites using comparative regional analysis". In: *Water Resources Research* 59.7, e2022WR033470. 10.1029/2022WR033470.

Haaf, E. and B. Heudorfer (2018). Groundwater dynamics indices (0.1): [data set]. 10.5281/zenodo.1486058.

Haaf, E., M. Giese, B. Heudorfer, K. Stahl, and R. Barthel (2020). "Physiographic and climatic controls on regional groundwater dynamics". In: *Water Resources Research* 56.10, e2019WR026545. 10.1029/2019WR026545.

Haehnel, P., T. C. Rasmussen, and G. C. Rau (2024). "Technical note: Removing dynamic sea-level influences from groundwater-level measurements". In: *Hydrology and Earth System Sciences* 28.12, 2767–2784. 10.5194/hess-28-2767-2024.

Hamlington, B. D., A. Bellas-Manley, J. K. Willis, S. Fournier, N. Vinogradova, R. S. Nerem, C. G. Piecuch, P. R. Thompson, and R. Kopp (2024). "The rate of global sea level rise doubled during the past three decades". In: *Communications Earth & Environment* 5, 601. 10.1038/s43247-024-01761-5.

Hannaford, J., G. Buys, K. Stahl, and L. M. Tallaksen (2013). "The influence of decadal-scale variability on trends in long European streamflow records". In: *Hydrology and Earth System Sciences* 17.7, 2717–2733. 10.5194/hess-17-2717-2013.

Hartmann, J. and N. Moosdorf (2012). "The new global lithological map database GLiM: A representation of rock properties at the Earth surface". In: *Geochemistry, Geophysics, Geosystems* 13.12. 10.1029/2012GC004370.

Helton, A. M., P. J. Dennedy–Frank, R. E. Emanuel, S. C. Neubauer, K. H. Adams, M. Ardon, L. Band, K. M. Befus, H. Borstlap, and J. A. Duberstein (2025). "Over, under, and through: Hydrologic connectivity and the future of coastal landscape salinization". In: *Water Resources Research* 61.7, e2024WR038720. 10.1029/2024WR038720.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz–Sabater, J. Nicolas, C. Peubey, R. Radu, and D. Schepers (2020). "The ERA5 global reanalysis". In: *Quarterly Journal of the Royal Meteorological Society* 146.730, 1999–2049. 10.1002/qj.3803.

Heudorfer, B., E. Haaf, K. Stahl, and R. Barthel (2019). "Index-based characterization and quantification of groundwater dynamics". In: *Water Resources Research* 55.7, 5575–5592. 10.1029/2018WR024418.

Heudorfer, B., H. V. Gupta, and R. Loritz (2025). "Are Deep Learning Models in Hydrology Entity Aware?" In: *Geophysical Research Letters* 52.6, e2024GL113036. 10.1029/2024GL113036.

Heudorfer, B., T. Liesch, and S. Broda (2024). "On the challenges of global entity-aware deep learning models for groundwater level prediction". In: *Hydrology and Earth System Sciences* 28.3, 525–543. 10.5194/hess-28-525-2024.

Hochreiter, S. and J. Schmidhuber (1997). "Long Short-Term Memory". In: *Neural computation* 9.8, 1735–1780. 10.1162/neco.1997.9.8.1735.

Hora, T., V. Srinivasan, and N. B. Basu (2019). "The groundwater recovery paradox in South India". In: *Geophysical Research Letters* 46.16, 9602–9611. 10.1029/2019GL083525.

Hrachowitz, M., H. H. Savenije, G. Blöschl, J. J. McDonnell, M. Sivapalan, J. W. Pomeroy, B. Arheimer, T. Blume, M. P. Clark, and U. Ehret (2013). "A decade of Predictions in Ungauged Basins (PUB)—a review". In: *Hydrological Sciences Journal* 58.6, 1198–1255. 10.1080/02626667.2013.80318.

Huggins, X., T. Gleeson, D. Serrano, S. Zipper, F. Jehn, M. M. Rohde, R. Abell, K. Vigerstol, and A. Hartmann (2023). "Overlooked risks and opportunities in groundwatersheds of the world's protected areas". In: *Nature Sustainability* 6.7, 855–864. 10.1038/s41893-023-01086-9.

Huscroft, J., T. Gleeson, J. Hartmann, and J. Börker (2018). "Compiling and mapping global permeability of the unconsolidated and consolidated Earth: GLobal HYdrogeology MaPS 2.0 (GLHYMPS 2.0)". In: *Geophysical Research Letters* 45.4, 1897–1904. 10.1002/2017GL075860.

IGRAC (2025). The Global Groundwater Monitoring Network (GGMN): [data set]. 10.58154/6Z0Y-DA34. URL: https://ggmn.un-igrac.org/.

IPCC (2021). Climate change 2021: The Physical Science Basis: Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. 10.1017/9781009157896.

Iturbide, M., J. M. Gutiérrez, L. M. Alves, J. Bedia, R. Cerezo-Mota, E. Cimadevilla, A. S. Cofiño, A. Di Luca, S. H. Faria, I. V. Gorodetskaya, M. Hauser, S. Herrera, K. Hennessy, H. T. Hewitt, R. G. Jones, S. Krakovska, R. Manzanas, D. Martínez-Castro, G. T. Narisma, I. S. Nurhati, I. Pinto, S. I. Seneviratne, B. van den Hurk, and C. S. Vera (2020). "An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets". In: *Earth System Science Data* 12.4, 2959–2970. 10.5194/essd-12-2959-2020.

Jakovovic, D., A. D. Werner, and C. T. Simmons (2011). "Numerical modelling of saltwater up-coning: comparison with experimental laboratory observations". In: *Journal of Hydrology* 402.3-4, 261–273. 10.1016/j.jhydrol.2011.03.021.

Jasechko, S., D. Perrone, H. Seybold, Y. Fan, and J. W. Kirchner (2020). "Groundwater level observations in 250,000 coastal US wells reveal scope of potential seawater intrusion". In: *Nature Communications* 11, 3229. 10.1038/s41467-020-17038-2.

Jasechko, S., H. Seybold, D. Perrone, Y. Fan, M. Shamsudduha, R. G. Taylor, O. Fallatah, and J. W. Kirchner (2024). "Rapid groundwater decline and some cases of recovery in aquifers globally". In: *Nature* 625.7996, 715–721. 10.1038/s41586-023-06879-8.

Johnson, T. D. and K. Belitz (2009). "Assigning land use to supply wells for the statistical characterization of regional groundwater quality: correlating urban land use and VOC occurrence". In: *Journal of Hydrology* 370.1-4, 100–108. 10.1016/j.jhydrol.2009.02.056.

Karger, D. N., O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, H. P. Linder, and M. Kessler (2017). "Climatologies at high resolution for the earth's land surface areas". In: *Scientific Data* 4, 1–20. 10.1038/sdata.2017.122.

Kaufman, L. and P. J. Rousseeuw (2009). Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

Kingma, D. P. and J. L. Ba (2014). "Adam: A method for stochastic optimization". In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. 10.48550/arxiv.1412.6980.

Kirchner, J. W. (2006). “Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology”. In: *Water Resources Research* 42.3. 10.1029/2005WR004362.

Klingler, C., K. Schulz, and M. Herrnegger (2021). “LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe”. In: *Earth System Science Data* 13.9, 4529–4565. 10.5194/essd-13-4529-2021.

Klotz, D., F. Kratzert, M. Gauch, A. Keefe Sampson, J. Brandstetter, G. Klambauer, S. Hochreiter, and G. Nearing (2022). “Uncertainty estimation with deep learning for rainfall–runoff modeling”. In: *Hydrology and Earth System Sciences* 26.6, 1673–1693. 10.5194/hess-26-1673-2022.

Knoll, L., L. Breuer, and M. Bach (2019). “Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning”. In: *Science of the Total Environment* 668, 1317–1327. 10.1016/j.scitotenv.2019.03.045.

Knowling, M. J., A. D. Werner, and D. Herckenrath (2015). “Quantifying climate and pumping contributions to aquifer depletion using a highly parameterised groundwater model: Uley South Basin (South Australia)”. In: *Journal of Hydrology* 523, 515–530. 10.1016/j.jhydrol.2015.01.081.

Konikow, L. F. (2011). “Contribution of global groundwater depletion since 1900 to sea–level rise”. In: *Geophysical Research Letters* 38.17. 10.1029/2011GL048604.

Kopp, R. E., M. Oppenheimer, J. L. O’Reilly, S. S. Drijfhout, T. L. Edwards, B. Fox-Kemper, G. G. Garner, N. R. Golledge, T. H. J. Hermans, and H. T. Hewitt (2023). “Communicating future sea-level rise uncertainty and ambiguity to assessment users”. In: *Nature Climate Change* 13.7, 648–660. 10.1038/s41558-023-01691-8.

Kratzert, F., M. Gauch, D. Klotz, and G. Nearing (2024). “HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin”. In: *Hydrology and Earth System Sciences* 28.17, 4187–4201. 10.5194/hess-28-4187-2024.

Kratzert, F., D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger (2018). “Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks”. In: *Hydrology and Earth System Sciences* 22.11, 6005–6022. 10.5194/hess-22-6005-2018.

Kratzert, F., D. Klotz, S. Hochreiter, and G. S. Nearing (2021). “A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall-runoff modeling”. In: *Hydrology and Earth System Sciences* 25.5, 2685–2703. 10.5194/hess-25-2685-2021.

Kratzert, F., D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing (2019). “Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets”. In: *Hydrology and Earth System Sciences* 23.12, 5089–5110. 10.5194/hess-23-5089-2019.

Kretschmer, D. V., H. A. Michael, N. Moosdorf, G. H. P. Oude Essink, M. F. P. Bierkens, T. Wagener, and R. Reinecke (2025a). “A perceptual model of drivers and limiters of coastal groundwater dynamics”. In: *Hydrological Processes* 39.1, e70058. 10.1002/hyp.70058.

Kretschmer, D. V., H. Michael, N. Moosdorf, G. O. Essink, M. F. P. Bierkens, T. Wagener, and R. Reinecke (2025b). "Controls on coastal saline groundwater across North America". In: *Environmental Research Letters* 20.2, 024065. 10.1088/1748-9326/ada973.

Kuang, X., J. Liu, B. R. Scanlon, J. J. Jiao, S. Jasechko, M. Lancia, B. K. Biskaborn, Y. Wada, H. Li, and Z. Zeng (2024). "The changing nature of groundwater in the global water cycle". In: *Science* 383.6686, eadf0630. 10.1126/science.adf0630.

Kuffer, M., M. Owusu, L. Oliveira, R. Sliuzas, and F. van Rijn (2022). "The missing millions in maps: Exploring causes of uncertainties in global gridded population datasets". In: *ISPRS International Journal of Geo-Information* 11.7, 403. 10.3390/ijgi11070403.

Kulp, S. A. and B. H. Strauss (2018). "CoastalDEM: A global coastal digital elevation model improved from SRTM using a neural network". In: *Remote Sensing of Environment* 206, 231–239. 10.1016/j.rse.2017.12.026.

Kulp, S. A. and B. H. Strauss (2019). "New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding". In: *Nature Communications* 10, 4844. 10.1038/s41467-019-12808-z.

Kumar, R., J. L. Musuuza, A. F. van Loon, A. J. Teuling, R. Barthel, J. ten Broek, J. Mai, L. Samaniego, and S. Attinger (2016). "Multiscale evaluation of the Standardized Precipitation Index as a groundwater drought indicator". In: *Hydrology and Earth System Sciences* 20.3, 1117–1131. 10.5194/hess-20-1117-2016.

Kunz, S., A. Schulz, M. Wetzel, M. Nölscher, T. Chiaburu, F. Biessmann, and S. Broda (2025). "Towards a global spatial machine learning model for seasonal groundwater level predictions in Germany". In: *Hydrology and Earth System Sciences* 29.15, 3405–3433. 10.5194/hess-29-3405-2025.

Kuss, A. J. M. and J. J. Gurdak (2014). "Groundwater level response in US principal aquifers to ENSO, NAO, PDO, and AMO". In: *Journal of Hydrology* 519, 1939–1952. 10.1016/j.jhydrol.2014.09.069.

Lall, U., L. Josset, and T. Russo (2020). "A Snapshot of the World's Groundwater Challenges". In: *Annual Review of Environment and Resources* 45, 171–194. 10.1146/annurev-environ-102017-025800.

Láng-Ritter, J., M. Keskinen, and H. Tenkanen (2025). "Global gridded population datasets systematically underrepresent rural population". In: *Nature Communications* 16, 2170. 10.1038/s41467-025-56906-7.

Langevin, C. D., J. D. Hughes, E. R. Banta, R. G. Niswonger, S. Panday, and A. M. Provost (2017). Documentation for the MODFLOW 6 groundwater flow model. Reston, Virginia: US Geological Survey.

Lecher, A. L. and K. R. M. Mackey (2018). "Synthesizing the effects of submarine groundwater discharge on marine biota". In: *Hydrology* 5.4, 60. 10.3390/hydrology5040060.

Lee, S., K.-K. Lee, and H. Yoon (2019). "Using artificial neural network models for groundwater level forecasting and assessment of the relative impacts of influencing factors". In: *Hydrogeology Journal* 27.2, 567–579. 10.1007/s10040-018-1866-3.

Lees, T., M. Buechel, B. Anderson, L. Slater, S. Reece, G. Coxon, and S. J. Dadson (2021). "Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models". In: *Hydrology and Earth System Sciences* 25.10, 5517–5534. 10.5194/hess-25-5517-2021.

Lehner, B. and G. Grill (2013). "Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems". In: *Hydrological Processes* 27.15, 2171–2186. 10.1002/hyp.9740.

Lehr, C. and G. Lischeid (2020). "Efficient screening of groundwater head monitoring data for anthropogenic effects and measurement errors". In: *Hydrology and Earth System Sciences* 24.2, 501–513. 10.5194/hess-24-501-2020.

LfU-SH (2003). Hydrogeologische Räume und Teilräume bezogen auf die oberflächennahen Wasserleiter, Landesamt für Umwelt des Landes Schleswig-Holstein. URL: https://umweltportal.schleswig-holstein.de/trefferanzeige?docuuid=d7a7934c-d125-4fc4-93cc-08f34a2a3d7c (visited on 03/30/2023).

Li, X., A. Khandelwal, X. Jia, K. Cutler, R. Ghosh, A. Renganathan, S. Xu, K. Tayal, J. Nieber, and C. Duffy (2022). "Regionalization in a global hydrologic deep learning model: from physical descriptors to random vectors". In: *Water Resources Research* 58.8, e2021WR031794. 10.1029/2021WR031794.

Linke, S., B. Lehner, C. Ouellet Dallaire, J. Ariwi, G. Grill, M. Anand, P. Beames, V. Burchard-Levine, S. Maxwell, H. Moidu, F. Tan, and M. Thieme (2019). "Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution". In: *Scientific Data* 6, 283. 10.1038/s41597-019-0300-6.

Lischeid, G., R. Dannowski, K. Kaiser, G. Nützmann, J. Steidl, and P. Stüve (2021). "Inconsistent hydrological trends do not necessarily imply spatially heterogeneous drivers". In: *Journal of Hydrology* 596, 126096. 10.1016/j.jhydrol.2021.126096.

Liu, P.-W., J. S. Famiglietti, A. J. Purdy, K. H. Adams, A. L. McEvoy, J. T. Reager, R. Bindlish, D. N. Wiese, C. H. David, and M. Rodell (2022a). "Groundwater depletion in California's Central Valley accelerates during megadrought". In: *Nature Communications* 13, 7825. 10.1038/s41467-022-35582-x.

Liu, Q., D. Gui, L. Zhang, J. Niu, H. Dai, G. Wei, and B. X. Hu (2022b). "Simulation of regional groundwater levels in arid regions using interpretable machine learning models". In: *Science of the Total Environment* 831, 154902. 10.1016/j.scitotenv.2022.154902.

Loritz, R., C. H. Wu, D. Klotz, M. Gauch, F. Kratzert, and M. Bassiouni (2024). "Generalizing tree–level sap flow across the European continent". In: *Geophysical Research Letters* 51.6, e2023GL107350. 10.1029/2023GL107350.

Luijendijk, E., T. Gleeson, and N. Moosdorf (2020). "Fresh groundwater discharge insignificant for the world's oceans but important for coastal ecosystems". In: *Nature Communications* 11, 1260. 10.1038/s41467-020-15064-8.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2, 56–67. 10.1038/s42256-019-0138-9.

Ma, K., D. Feng, K. Lawson, W.-.-P. Tsai, C. Liang, X. Huang, A. Sharma, and C. Shen (2021). "Transferring hydrologic data across continents–leveraging data–rich regions to improve hydrologic prediction in data–sparse regions". In: *Water Resources Research* 57.5, e2020WR028600. 10.1029/2020WR028600.

Ma, Y., E. Leonarduzzi, A. Defnet, P. Melchior, L. E. Condon, and R. M. Maxwell (2024). "Water table depth estimates over the contiguous united states using a random forest model". In: *Groundwater* 62, 34–43. 10.1111/gwat.13362.

Mádl-Szőnyi, J., O. Batelaan, J. Molson, H. Verweij, X.-W. Jiang, J. J. Carrillo-Rivera, and Á. Tóth (2023). "Regional groundwater flow and the future of hydrogeology: evolving concepts and communication". In: *Hydrogeology Journal* 31, 23–26. 10.1007/s10040-022-02577-3.

Mangor, K., N. K. Drønen, K. H. Kærgaard, and S. E. Kristensen (2017). Shoreline management guidelines. Hørsholm, Denmark: DHI.

Mangukiya, N. K. and A. Sharma (2025). "Deep learning–based approach for enhancing streamflow prediction in watersheds with aggregated and intermittent observations". In: *Water Resources Research* 61, e2024WR037331. 10.1029/2024WR037331.

Martens, K., M. van Camp, D. van Damme, and K. Walraevens (2013). "Groundwater dynamics converted to a groundwater classification as a tool for nature development programs in the dunes". In: *Journal of Hydrology* 499, 236–246. 10.1016/j.jhydrol.2013.06.045.

Martínez, M. L., A. Intralawan, G. Vázquez, O. Pérez-Maqueo, P. Sutton, and R. Landgrave (2007). "The coasts of our world: Ecological, economic and social importance". In: *Ecological Economics* 63.2-3, 254–272. 10.1016/j.ecolecon.2006.10.022.

Massei, N., D. G. Kingston, D. M. Hannah, J.-P. Vidal, B. Dieppois, M. Fossa, A. Hartmann, D. A. Lavers, and B. Laignel (2020). "Understanding and predicting large-scale hydrological variability in a changing environment". In: *Proceedings of the International Association of Hydrological Sciences* 383, 141–149. 10.5194/piahs-383-141-2020.

McKee, T. B., N. J. Doesken, and J. Kleist (1993). "The relationship of drought frequency and duration to time scales". In: *Proceedings of the 8th Conference on Applied Climatology*, 179–183.

McMillan, H. (2020). "Linking hydrologic signatures to hydrologic processes: A review". In: *Hydrological Processes* 34.6, 1393–1409. 10.1002/hyp.13632.

Mekanik, F., M. A. Imteaz, S. Gato-Trinidad, and A. Elmahdi (2013). "Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes". In: *Journal of Hydrology* 503, 11–21. 10.1016/j.jhydrol.2013.08.035.

Michael, H. A., A. E. Mulligan, and C. F. Harvey (2005). "Seasonal oscillations in water exchange between aquifers and the coastal ocean". In: *Nature* 436.7054, 1145–1148. 10.1038/nature03935.

Michael, H. A., C. J. Russoniello, and L. A. Byron (2013). "Global assessment of vulnerability to sea–level rise in topography–limited and recharge–limited coastal groundwater systems". In: *Water Resources Research* 49.4, 2228–2240. 10.1002/wrcr.20213.

Mishra, N., D. Khare, K. K. Gupta, and R. Shukla (2014). "Impact of land use change on groundwater - a review". In: *Advances in Water Resource and Protection* 2.28, 28–41.

Moeck, C., N. Grech-Cumbo, J. Podgorski, A. Bretzler, J. J. Gurdak, M. Berg, and M. Schirmer (2020). "A global-scale dataset of direct natural groundwater recharge rates: A review of variables, processes and relationships". In: *Science of the Total Environment* 717, 137042. 10.1016/j.scitotenv.2020.137042.

Mohammadi, M., S. Gato-Trinidad, and K. King Kuok (2024). "The limitation of machine learning methods for water supply and demand forecasting: A case study for Greater Melbourne, Australia". In: *Water Supply* 24.11, 3848–3861. 10.2166/ws.2024.225.

Moosdorf, N. and T. Oehler (2017). "Societal use of fresh submarine groundwater discharge: An overlooked water resource". In: *Earth-Science Reviews* 171, 338–348. 10.1016/j.earscirev.2017.06.006.

Morgan, L. K. and A. D. Werner (2015). "A national inventory of seawater intrusion vulnerability for Australia". In: *Journal of Hydrology: Regional Studies* 4, 686–698. 10.1016/j.ejrh.2015.10.005.

Mueller, W., D. Zamrsky, G. O. Essink, L. E. Fleming, A. Deshpande, K. C. Makris, B. W. Wheeler, J. N. Newton, K. V. Narayan, and A. M. Naser (2024). "Saltwater intrusion and human health risks for coastal populations under 2050 climate scenarios". In: *Scientific Reports* 14, 15881. 10.1038/s41598-024-66956-4.

Müller Schmied, H., D. Cáceres, S. Eisner, M. Flörke, C. Herbert, C. Niemann, T. A. Peiris, E. Popat, F. T. Portmann, R. Reinecke, S. Shadkam, T. Trautmann, and P. Döll (2020). The global water resources and use model WaterGAP v2.2d - Standard model output: [data set]. 10.1594/PANGAEA.918447.

Najah, A., A. El-Shafie, O. A. Karim, and A. H. El-Shafie (2013). "Application of artificial neural networks for water quality prediction". In: *Neural Computing and Applications* 22, 187–201. 10.1007/s00521-012-0940-3.

Narvaez-Montoya, C., J. Mahlknecht, J. A. Torres-Martínez, A. Mora, and G. Bertrand (2023). "Seawater intrusion pattern recognition supported by unsupervised learning: A systematic review and application". In: *Science of the Total Environment* 864, 160933. 10.1016/j.scitotenv.2022.160933.

NASA JPL (2020). NASADEM Merged DEM Global 1 arc second V001: [data set]. 10.5067 /MEASURES/NASADEM/NASADEM_HGT.001.

Nazari, S., R. Reinecke, and N. Moosdorf (2025). "Global estimates of groundwater withdrawal trends and uncertainties". In: *Environmental Research Letters* 20.9, 094043. 10.1088/1748-93 26/adf6ca.

Nearing, G., D. Cohen, V. Dube, M. Gauch, O. Gilon, S. Harrigan, A. Hassidim, D. Klotz, F. Kratzert, and A. Metzger (2024). "Global prediction of extreme floods in ungauged watersheds". In: *Nature* 627.8004, 559–563. 10.1038/s41586-024-07145-1.

Neuman, S. P. and V. Di Federico (2003). "Multifaceted nature of hydrogeologic scaling and its interpretation". In: *Reviews of Geophysics* 41.3. 10.1029/2003RG000130.

Neumann, B., A. T. Vafeidis, J. Zimmermann, and R. J. Nicholls (2015). "Future coastal population growth and exposure to sea-level rise and coastal flooding-a global assessment". In: *PloS one* 10.3, e0118571. 10.1371/journal.pone.0118571.

Nimmo, J. R., K. S. Perkins, M. R. Plampin, M. A. Walvoord, B. A. Ebel, and B. B. Mirus (2021). "Rapid-response unsaturated zone hydrology: Small-scale data, small-scale theory, big problems". In: *Frontiers in Earth Science* 9, 613564. 10.3389/feart.2021.613564.

Nölscher, M., M. Mutz, and S. Broda (2022). "Multiorder hydrologic position for Europe - a set of features for machine learning and analysis in hydrology". In: *Scientific Data* 9, 662. 10.1038/s41597-022-01787-4.

Nolte, **A.** (2026). Coastal GroundWater Level (CGWL) Dataset: [data set]. 10.5281/zenodo.15 008065.

Nolte, **A.**, S. Bender, J. Hartmann, S. Baltruschat, N. Moosdorf, and R. Reinecke (2026). "Coastal groundwater-level trends reveal global susceptibility to seawater intrusion". In: *Nature Water* 4, 673–682. 10.1038/s44221-026-00619-8.

Nolte, **A.**, M. Eley, M. Schöniger, D. Gwapedza, J. Tanner, S. K. Mantel, and K. Scheihing (2021). "Hydrological modelling for assessing spatio–temporal groundwater recharge variations in the water–stressed Amathole Water Supply System, Eastern Cape, South Africa". In: *Hydrological Processes* 35.6, e14264. 10.1002/hyp.14264.

Nolte, **A.**, E. Haaf, B. Heudorfer, S. Bender, and J. Hartmann (2024). "Disentangling coastal groundwater level dynamics in a global dataset". In: *Hydrology and Earth System Sciences* 28.5, 1215–1249. 10.5194/hess-28-1215-2024.

Nolte, **A.**, B. Heudorfer, S. Bender, and J. Hartmann (2025). "Multi-site deep learning for groundwater level prediction across global datasets: toward scalable applications under data scarcity". In: *Journal of Hydroinformatics* 27.10, 1632–1651. 10.2166/hydro.2025.095.

Olden, J. D., M. J. Kennard, and B. J. Pusey (2012). "A framework for hydrologic classification with a review of methodologies and applications in ecohydrology". In: *Ecohydrology* 5.4, 503–518. 10.1002/eco.251.

Ostrom, E. (1992). Governing the commons: The evolution of institutions for collective action. Cambridge University Press.

Otto, R. (2001). “Estimating groundwater recharge rates in the southeastern Holstein region, northern Germany”. In: *Hydrogeology Journal* 9.5, 498–511. 10.1007/s100400100155.

Oude Essink, G. H. P., E. S. van Baaren, and P. G. B. de Louw (2010). “Effects of climate change on coastal groundwater systems: A modeling study in the Netherlands”. In: *Water Resources Research* 46.10. 10.1029/2009WR008719.

Panthi, J., S. M. Pradhanang, **A. Nolte**, and T. B. Boving (2022). “Saltwater intrusion into coastal aquifers in the contiguous United States - A systematic review of investigation approaches and monitoring networks”. In: *Science of the Total Environment* 836, 155641. 10.1016/j.scitotenv.2022.155641.

Papacharalampous, G. and H. Tyralis (2022). “Time series features for supporting hydrometeorological explorations and predictions in ungauged locations using large datasets”. In: *Water* 14.10, 1657. 10.3390/w14101657.

Papacharalampous, G., H. Tyralis, Y. Markonis, and M. Hanel (2023). “Hydroclimatic time series features at multiple time scales”. In: *Journal of Hydrology* 618, 129160. 10.1016/j.jhydrol.2023.129160.

Parisi, A., M. R. Alfio, G. Balacco, C. Güler, and M. D. Fidelibus (2023). “Analyzing spatial and temporal evolution of groundwater salinization through Multivariate Statistical Analysis and Hydrogeochemical Facies Evolution-Diagram”. In: *Science of the Total Environment* 862, 160697. 10.1016/j.scitotenv.2022.160697.

Peng, J., A. Loew, O. Merlin, and N. E. C. Verhoest (2017). “A review of spatial downscaling of satellite remotely sensed soil moisture”. In: *Reviews of Geophysics* 55.2, 341–366. 10.1002/2016RG000543.

Perrone, D. and S. Jasechko (2019). “Deeper well drilling an unsustainable stopgap to groundwater depletion”. In: *Nature Sustainability* 2, 773–782. 10.1038/s41893-019-0325-z.

Peters, C. N., C. Kimsal, R. S. Frederiks, A. Paldor, R. McQuiggan, and H. A. Michael (2022). “Groundwater pumping causes salinization of coastal streams due to baseflow depletion: Analytical framework and application to Savannah River, GA”. In: *Journal of Hydrology* 604, 127238. 10.1016/j.jhydrol.2021.127238.

Poggio, L., L. M. de Sousa, N. H. Batjes, G. B. M. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter (2021). “SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty”. In: *Soil* 7.1, 217–240. 10.5194/soil-7-217-2021.

Post, V., H. Kooi, and C. Simmons (2007). “Using hydraulic head measurements in variable–density ground water flow analyses”. In: *Groundwater* 45.6, 664–671. 10.1111/j.1745-6584.2007.00339.x.

Probst, P., M. N. Wright, and A.-.-L. Boulesteix (2019). “Hyperparameters and tuning strategies for random forest”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.3, e1301. 10.1002/widm.1301.

Python Software Foundation (2021). Python language reference, version 3.7.11: [software]. URL: https://www.python.org.

R Core Team (2021). R: A language and environment for statistical computing: [software]. URL: https://www.R-project.org/.

Rajaee, T., H. Ebrahimi, and V. Nourani (2019). "A review of the artificial intelligence methods in groundwater level modeling". In: *Journal of Hydrology* 572, 336–351. 10.1016/j.jhydrol.2018.12.037.

Rasmussen, P., T. O. Sonnenborg, G. Goncear, and K. Hinsby (2013). "Assessing impacts of climate change, sea level rise, and drainage canals on saltwater intrusion to coastal aquifer". In: *Hydrology and Earth System Sciences* 17, 421–443. 10.5194/hess-17-421-2013.

Rau, G. C., M. O. Cuthbert, V. E. A. Post, D. Schweizer, R. I. Acworth, M. S. Andersen, P. Blum, E. Carrara, T. C. Rasmussen, and S. Ge (2020). "Future-proofing hydrogeology by revising groundwater monitoring practice". In: *Hydrogeology Journal* 28.8, 2963–2969. 10.1007/s10040-020-02242-7.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat (2019). "Deep learning and process understanding for data-driven Earth system science". In: *Nature* 566.7743, 195–204. 10.1038/s41586-019-0912-1.

Reinecke, R., L. Foglia, S. Mehl, T. Trautmann, D. Cáceres, and P. Döll (2019). "Challenges in developing a global gradient-based groundwater model (G3M v1.0) for the integration into a global hydrological model". In: *Geoscientific Model Development* 12.6, 2401–2418. 10.5194/gmd-12-2401-2019.

Reinecke, R., S. Gnann, L. Stein, M. Bierkens, I. de Graaf, T. Gleeson, G. O. Essink, E. H. Sutanudjaja, C. R. Vargas, and J. Verkaik (2024). "Uncertainty in model estimates of global groundwater depth". In: *Environmental Research Letters* 19.11, 114066. 10.1088/1748-9326/ad8587.

Reinecke, R., H. Müller Schmied, T. Trautmann, L. S. Andersen, P. Burek, M. Flörke, S. N. Gosling, M. Grillakis, N. Hanasaki, and A. Koutroulis (2021). "Uncertainty of simulated groundwater recharge at different global warming levels: a global-scale multi-model ensemble study". In: *Hydrology and Earth System Sciences* 25.2, 787–810. 10.5194/hess-25-787-2021.

Retike, I., J. Bikše, A. Kalvāns, A. Dēliņa, Z. Avotniece, W. J. Zaadnoordijk, M. Jemeljanova, K. Popovs, A. Babre, and A. Zelenkevičs (2022). "Rescue of groundwater level time series: How to visually identify and treat errors". In: *Journal of Hydrology* 605, 127294. 10.1016/j.jhydrol.2021.127294.

Richardson, C. M., K. L. Davis, C. Ruiz-González, J. A. Guimond, H. A. Michael, A. Paldor, N. Moosdorf, and A. Paytan (2024). "The impacts of climate change on coastal groundwater". In: *Nature Reviews Earth & Environment* 5, 100–119. 10.1038/s43017-023-00500-2.

Richts, A., W. F. Struckmeier, and M. Zaepke (2011). WHYMAP and the Groundwater Resources Map of the World 1:25,000,000. Dordrecht: Springer. 10.1007/978-90-481-3426-7_10.

Riedel, T. and T. K. D. Weber (2020). "Review: The influence of global change on Europe's water cycle and groundwater recharge". In: *Hydrogeology Journal* 28.6, 1939–1959. 10.1007/s10040-020-02165-3.

Rinderer, M., H. J. Meerveld, and B. L. McGlynn (2019). "From points to patterns: using groundwater time series clustering to investigate subsurface hydrological connectivity and runoff source area dynamics". In: *Water Resources Research* 55.7, 5784–5806. 10.1029/2018WR023886.

Rodriguez, E., C. S. Morris, and J. E. Belz (2006). "A global assessment of the SRTM performance". In: *Photogrammetric Engineering & Remote Sensing* 72.3, 249–260. 10.14358/PERS.72.3.249.

Rousseeuw, P. J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, 53–65. 10.1016/0377-0427(87)90125-7.

Sanford, W. (2002). "Recharge and groundwater models: an overview". In: *Hydrogeology Journal* 10, 110–120. 10.1007/s10040-001-0173-5.

Sayre, R., K. Butler, K. van Graafeiland, S. Breyer, D. Wright, C. Frye, D. Karagulle, M. Martin, J. Cress, T. Allen, R. Allee, R. Parsons, B. Nyberg, M. Costello, P. Harris, and F. Muller-Karger (2021). "A global ecological classification of coastal segment units to complement marine biodiversity observation network assessments". In: *Oceanography* 34.2, 120–129. 10.5670/oceanog.2021.219.

Scanlon, B. R., R. W. Healy, and P. G. Cook (2002). "Choosing appropriate techniques for quantifying groundwater recharge". In: *Hydrogeology Journal* 10, 18–39. 10.1007/s10040-001-0176-2.

Serinaldi, F., C. G. Kilsby, and F. Lombardo (2018). "Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology". In: *Advances in Water Resources* 111, 132–155. 10.1016/j.advwatres.2017.10.015.

Shangguan, W., T. Hengl, J. Mendes de Jesus, H. Yuan, and Y. Dai (2017). "Mapping the global depth to bedrock for land surface modeling". In: *Journal of Advances in Modeling Earth Systems* 9.1, 65–88. 10.1002/2016MS000686.

Shen, C., E. Laloy, A. Elshorbagy, A. Albert, J. Bales, F.-J. Chang, S. Ganguly, K.-L. Hsu, D. Kifer, Z. Fang, K. Fang, D. Li, X. Li, and W.-P. Tsai (2018). "HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community". In: *Hydrology and Earth System Sciences* 22.11, 5639–5656. 10.5194/hess-22-5639-2018.

Shen, C. and K. Lawson (2021). "Applications of deep learning in hydrology". In: *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, 283–297. 10.1002/9781119646181.ch19.

Shokri–Kuehni, S. M. S., B. Raaijmakers, T. Kurz, D. Or, R. Helmig, and N. Shokri (2020). "Water table depth and soil salinization: From pore–scale processes to field–scale responses". In: *Water Resources Research* 56.2, e2019WR026707. 10.1029/2019WR026707.

Smerdon, B. D. (2017). "A synopsis of climate change effects on groundwater recharge". In: *Journal of Hydrology* 555, 125–128. 10.1016/j.jhydrol.2017.09.047.

Solgi, R., H. A. Loáiciga, and M. Kram (2021). "Long short-term memory neural network (LSTM-NN) for aquifer level time series forecasting using in-situ piezometric observations". In: *Journal of Hydrology* 601, 126800. 10.1016/j.jhydrol.2021.126800.

Sophocleous, M. (2002). "Interactions between groundwater and surface water: the state of the science". In: *Hydrogeology Journal* 10, 52–67. 10.1007/s10040-001-0170-8.

Sorensen, J. P. R., J. Davies, G. Y. Ebrahim, J. Lindle, B. P. Marchant, M. J. Ascott, J. P. Bloomfield, M. O. Cuthbert, M. Holland, K. H. Jensen, M. Shamsudduha, K. G. Villholth, A. M. MacDonald, and R. G. Taylor (2021). "The influence of groundwater abstraction on interpreting climate controls and extreme recharge events from well hydrographs in semi-arid South Africa". In: *Hydrogeology Journal* 29.8, 2773–2787. 10.1007/s10040-021-02391-3.

Spinoni, J., P. Barbosa, M. Cherlet, G. Forzieri, N. McCormick, G. Naumann, J. V. Vogt, and A. Dosio (2021). "How will the progressive global increase of arid areas affect population and land-use in the 21st century?" In: *Global and Planetary Change* 205, 103597. 10.1016/j.glopl acha.2021.103597.

Strehl, A. and J. Ghosh (2002). "Cluster ensembles - a knowledge reuse framework for combining multiple partitions". In: *Journal of Machine Learning Research* 3, 583–617.

Sutanudjaja, E. H., L. P. van Beek, S. M. de Jong, F. C. van Geer, and M. F. Bierkens (2011). "Large-scale groundwater modeling using global datasets: a test case for the Rhine-Meuse basin". In: *Hydrology and Earth System Sciences* 15.9, 2913–2935. 10.5194/hess-15-2913-2011.

Taniguchi, M., W. C. Burnett, J. E. Cable, and J. V. Turner (2002). "Investigation of submarine groundwater discharge". In: *Hydrological Processes* 16.11, 2115–2129. 10.1002/hyp.1145.

Tao, H., M. M. Hameed, H. A. Marhoon, M. Zounemat-Kermani, S. Heddam, S. Kim, S. O. Sulaiman, M. L. Tan, Z. Sa'adi, and A. D. Mehr (2022). "Groundwater level prediction using machine learning models: A comprehensive review". In: *Neurocomputing* 489, 271–308. 10.1016/j.neucom.2022.03.014.

Taylor, C. J. and W. M. Alley (2001). Ground-water-level monitoring and the importance of long-term water-level data. Vol. 1217. U.S. Geological Circular. Denver, USA: U.S. Geological Survey.

Taylor, R. G., B. Scanlon, P. Döll, M. Rodell, R. van Beek, Y. Wada, L. Longuevergne, M. Leblanc, J. S. Famiglietti, and M. Edmunds (2013). "Ground water and climate change". In: *Nature Climate Change* 3.4, 322–329. 10.1038/nclimate1744.

Thorslund, J. and M. T. H. van Vliet (2020). "A global dataset of surface water and groundwater salinity measurements from 1980-2019". In: *Scientific Data* 7, 231. 10.1038/s41597-020-0562-z.

Trabucco, A. and R. Zomer (2019). Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2: [data set]. 10.6084/m9.figshare.7504448.v3.

Tully, K., K. Gedan, R. Epanchin-Niell, A. Strong, E. S. Bernhardt, T. BenDor, M. Mitchell, J. Kominoski, T. E. Jordan, and S. C. Neubauer (2019). "The invisible flood: The chemistry, ecology, and social implications of coastal saltwater intrusion". In: *BioScience* 69.5, 368–378. 10.1093/biosci/biz027.

Tyralis, H., G. Papacharalampous, and A. Langousis (2019). "A brief review of random forests for water scientists and practitioners and their recent history in water resources". In: *Water* 11.5, 910. 10.3390/w11050910.

Uber Technologies Inc (2024). H3: Hexagonal hierarchical geospatial indexing system: [software]. URL: https://h3geo.org/.

United Nations (2022). "The United Nations World Water Development Report 2022: Groundwater - Making the invisible visible". In: 10.18356/9789210015363.

Vahdat-Aboueshagh, H., F. T.-C. Tsai, D. Bhatta, and K. P. Paudel (2021). "Irrigation-intensive groundwater modeling of complex aquifer systems through integration of big geological data". In: *Frontiers in Water* 3. 10.3389/frwa.2021.623476.

van der Maaten, L. and G. Hinton (2008). "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9, 2579–2605.

van Loon, A. F., R. Kumar, and V. Mishra (2017). "Testing the use of standardised indices and GRACE satellite data to estimate the European 2015 groundwater drought in near-real time". In: *Hydrology and Earth System Sciences* 21.4, 1947–1971. 10.5194/hess-21-1947-2017.

Venâncio, C., R. Ribeiro, and I. Lopes (2022). "Seawater intrusion: An appraisal of taxa at most risk and safe salinity levels". In: *Biological Reviews* 97.1, 361–382. 10.1111/brv.12803.

Vicente-Serrano, S. M., S. Beguería, and J. I. López-Moreno (2010). "A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index". In: *Journal of Climate* 23.7, 1696–1718. 10.1175/2009JCLI2909.1.

Vries, J. J. de and I. Simmers (2002). "Groundwater recharge: an overview of processes and challenges". In: *Hydrogeology Journal* 10, 5–17. 10.1007/s10040-001-0171-7.

Wagenaar, D., S. Lüdtke, K. Schröter, L. M. Bouwer, and H. Kreibich (2018). "Regional and temporal transferability of multivariable flood damage models". In: *Water Resources Research* 54.5, 3688–3703. 10.1029/2017WR022233.

Wang, M., J. Yao, H. Chang, R. Liu, N. Xu, Z. Liu, H. Gong, H. Zheng, J. Wang, and X. Guo (2025). "Underground well water level observation grid dataset from 2005 to 2022". In: *Scientific Data* 12, 728. 10.1038/s41597-025-04799-y.

Wang, S., W. Zhou, and C. Jiang (2020). "A survey of word embeddings based on deep learning". In: *Computing* 102.3, 717–740. 10.1007/s00607-019-00768-7.

Wang, X., K. Smith, and R. Hyndman (2006). "Characteristic-based clustering for time series data". In: *Data mining and knowledge Discovery* 13.3, 335–364. 10.1007/s10618-005-0039-x.

Werner, A. D. (2017). "On the classification of seawater intrusion". In: *Journal of Hydrology* 551, 619–631. 10.1016/j.jhydrol.2016.12.012.

Werner, A. D., M. Bakker, V. E. A. Post, A. Vandenbohede, C. Lu, B. Ataie-Ashtiani, C. T. Simmons, and D. A. Barry (2013). "Seawater intrusion processes, investigation and management: recent advances and future challenges". In: *Advances in Water Resources* 51, 3–26. 10.1016/j.advwatres.2012.03.004.

Werner, A. D., J. D. Ward, L. K. Morgan, C. T. Simmons, N. I. Robinson, and M. D. Teubner (2012). "Vulnerability indicators of sea water intrusion". In: *Groundwater* 50, 48–58. 10.1111/j.1745-6584.2011.00817.x.

Wi, S. and S. Steinschneider (2024). "On the need for physical constraints in deep learning rainfall–runoff projections under climate change: a sensitivity analysis to warming and shifts in potential evapotranspiration". In: *Hydrology and Earth System Sciences* 28.3, 479–503. 10.5194/hess-28-479-2024.

Winkler, K., R. Fuchs, M. D. A. Rounsevell, and M. Herold (2020). HILDA+ Global Land Use Change between 1960 and 2019: [data set]. 10.1594/PANGAEA.921846.

Worland, S. C., S. Steinschneider, W. Asquith, R. Knight, and M. Wieczorek (2019). "Prediction and inference of flow duration curves using multioutput neural networks". In: *Water Resources Research* 55.8, 6850–6868. 10.1029/2018WR024463.

Wriedt, G. (2017). "Verfahren zur Analyse klimatischer und anthropogener Einflüsse auf die Grundwasserstandsentwicklung". In: *Grundwasser* 22, 41–53. 10.1007/s00767-016-0349-5.

Wunsch, A., T. Liesch, and S. Broda (2021a). "Feature-based groundwater hydrograph clustering using unsupervised self-organizing map-ensembles". In: *Water Resources Management* 36, 39–54. 10.1007/s11269-021-03006-y.

Wunsch, A., T. Liesch, and S. Broda (2021b). "Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX)". In: *Hydrology and Earth System Sciences* 25.3, 1671–1687. 10.5194/hess-25-1671-2021.

Wunsch, A., T. Liesch, and S. Broda (2022). "Deep learning shows declining groundwater levels in Germany until 2100 due to climate change". In: *Nature Communications* 13, 1221. 10.1038/s41467-022-28770-2.

Wunsch, A., T. Liesch, and N. Goldscheider (2023). "Towards understanding the influence of seasons on low groundwater periods based on explainable machine learning". In: *Hydrology and Earth System Sciences* 28.9, 2167–2178. 10.5194/hess-28-2167-2024.

Xanke, J. and T. Liesch (2022). "Quantification and possible causes of declining groundwater resources in the Euro-Mediterranean region from 2003 to 2020". In: *Hydrogeology Journal* 30.2, 379–400. 10.1007/s10040-021-02448-3.

Yang, Y. and T. F. M. Chui (2021). "Modeling and interpreting hydrological responses of sustainable urban drainage systems with explainable machine learning methods". In: *Hydrology and Earth System Sciences* 25.11, 5839–5858. 10.5194/hess-25-5839-2021.

Zamrsky, D., G. H. P. Oude Essink, and M. F. P. Bierkens (2018). "Estimating the thickness of unconsolidated coastal aquifers along the global coastline". In: *Earth System Science Data* 10.3, 1591–1603. 10.5194/essd-10-1591-2018.

Zamrsky, D., G. H. P. Oude Essink, and M. F. P. Bierkens (2024). "Global impact of sea level rise on coastal fresh groundwater resources". In: *Earth's Future* 12, e2023EF003581. 10.1029/2023EF003581.

Zhang, Y., C.-.-Y. Wang, L.-.-Y. Fu, and Q.-.-Y. Yang (2021). “Are deep aquifers really confined? Insights from deep groundwater tidal responses in the North China Platform”. In: *Water Resources Research* 57.11, e2021WR030195. 10.1029/2021WR030195.

Zomer, R. J., J. Xu, and A. Trabucco (2022). “Version 3 of the global aridity index and potential evapotranspiration database”. In: *Scientific Data* 9, 409. 10.1038/s41597-022-01493-1.

Appendix A The Coastal GroundWater Level (CGWL) dataset

A.1 Data acquisition and preprocessing

Each data source (Table 1) went through the following general preprocessing steps for data selection and standardization, conducted using Python (Python Software Foundation 2021) and ArcGIS Pro (Esri 2022):

a) Check for necessary metadata (in convertible units for homogenization):
   - Latitude and longitude of the well location.
   - Reference vertical point or datum used to report the GWL data.
   - Date of observation (exceptions: for source number 8 (Table 1), the given date for a measurement period was accepted; for annually aggregated data from the Jasechko et al. (2020) dataset included in source 14, the provided year was accepted as the observation date).

b) Optional secondary information:
   - Where available, commonly reported secondary data fields were standardized across datasets. Secondary data refer to attributes that provide context for interpreting GWL observations, such as aquifer state (e.g., confined, unconfined), geological composition (e.g., sandstone, limestone), and technical well specifications (e.g., casing type or intended use).

c) GWL time series filtering:
   - Removal of common hydrological placeholders for missing values (e.g., 999, -9999).
   - Time series with only zero values were removed, due to ambiguity in whether they represent valid or missing data.

d) Individual datasets were acquired and preprocessed with a preference for referencing GWLs to the ground surface rather than to sea level or the top of casing. Accordingly, all GWL observations were harmonized to a ground-surface reference (i.e., groundwater depth/WTD):

  - If GWLs were provided relative to sea level, they were converted to ground surface reference using (1) the elevation data included in the source dataset, or, if unavailable, (2) digital elevation models (DEMs). Specifically, CoastalDEM (Kulp and Strauss 2019) data were used where available; for more inland regions, SRTM data (NASA JPL 2020) were used. To assess the uncertainty in reference-point conversions, Figure A1 compares (1) with (2) showing a mean deviation of 3.11 m.
  - GWLs referenced to the top of casing were converted to ground surface using the respective wellhead elevation data, where available. Remaining measurements referenced to the top of casing were then also treated as depths relative to a local ground-surface proxy, which may add at most about 1 m uncertainty to WTD at individual wells.

e) Groundwater elevation time series were computed as surface elevation minus standardized groundwater depth, using either source-provided elevations (1) or DEM-derived elevations (2) (see d)).

f) Well location and situation:

  - Selection of wells located within 100 km distance of the coastline (coastline dataset: Lehner and Grill 2013).
  - Selection of shallow wells, following Fan et al. (2013): wells with median GWLs between -2 and 100 m below the reference point (mbrp).

g) Duplicate time series were removed by grouping wells with identical coordinates (latitude and longitude rounded to three decimals) and identical mean GWL (rounded to two decimals), retaining only the time series with the largest number of observations in each group.

In addition to these general steps, source-specific preprocessing was applied to all datasets (Section A.6), as metadata structures, data formats, and reporting conventions varied across sources and required tailored handling to ensure consistency.

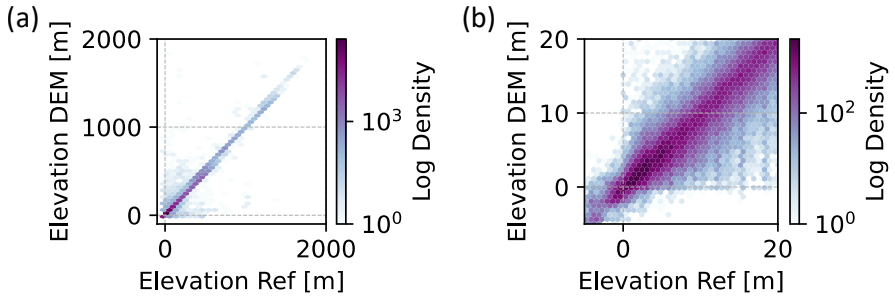


Figure A1: Comparison between well land-surface elevations provided in the groundwater dataset (Elevation Ref) and those derived from digital elevation models (Elevation DEM; see d)). The left panel shows all wells across the full elevation range, while the right panel provides a zoomed-in view of low-lying coastal areas. Elevation data from the two sources are strongly correlated (Pearson r = 0.99), with a mean absolute error (MAE) of 3.11 m across all wells. MAE is lower in low-lying areas (1.72 m for Elevation Ref $\leq$10 m) and higher in more elevated terrain (3.63 m for Elevation Ref >10 m). Overall, 50% of wells show an elevation difference greater than 1 m. Kulp and Strauss (2019) found that SRTM data exhibited a global mean vertical bias of 1.88 m, which CoastalDEM reduced to a bias of -0.29 m. In our analysis, larger differences – particularly in more inland regions – may reflect that SRTM data (NASA JPL 2020) were used in inland regions not covered by CoastalDEM, the use of the 900 m resolution open-source CoastalDEM product, and temporal inconsistencies in Elevation Ref across datasets, with variations of several meters found in some datasets (e.g., source number 21 in Table 1).

A.2 Limitations in differentiating aquifers at large scales

While the CGWL dataset is restricted to wells with median groundwater depths shallower than 100 m to favor systems that are at least partly connected to local recharge and hydroclimatic variability (Section A.1), measured WTD is fundamentally an unconfined concept. In confined systems, wells measure hydraulic head (potentiometric surface), which can rise above the top of the aquifer and even above the land surface (artesian). Well depth alone, therefore, does not allow inference about confinement. Even deep aquifers overlain by thick barrier formations can be leaky, and confinement may show no systematic depth dependence (Zhang et al. 2021). More generally, groundwater flow systems are three-dimensional. Groundwater basins encompass full flow paths, and hydraulic continuity can extend across multiple aquifers, intervening aquitards, and structural features such as faults (Mádl-Szőnyi et al. 2023).

Against this background, differentiating aquifers in the CGWL dataset is limited in two main ways. First, aquifer information for in situ GWL observations in the CGWL dataset is sparse. Aquifer state (e.g., confined, unconfined) is reported for only ~7% of wells, mainly from the United States, Denmark, Belgium, France, and Sweden (see Section A.6 for source-specific details). In addition, the hydrogeologic unit tapped by a given well is

rarely specified, so the monitored interval within a multi-layered aquifer system generally cannot be determined. Second, currently available global hydrogeologic datasets primarily describe near-surface properties and do not resolve vertical aquifer layering or confining units. For example, the unconsolidated thickness map of Zamrsky et al. (2018) provides an estimate of sedimentary cover thickness along coasts, and GLHYMPS (Huscroft et al. 2018) provides lithology-based permeability, including mapped unconsolidated surface sediments. However, these layers do not necessarily reflect the aquifer containing the well since high-permeability unconsolidated cover can overlie deeper, potentially confined storeys in sedimentary basins and alluvial plains, which commonly comprise stacked aquifers and aquitards. Comparing the wells with reported aquifer state (7%) in the CGWL dataset with GLHYMPS-based permeability distributions in Figure A2 also illustrates the difficulty of deriving permeability thresholds for different aquifer states at the global scale. In principle, combining different information on near-surface hydrogeology could, in the future, help to better constrain where unconfined conditions are likely near the surface. However, to our knowledge, such approaches have not yet been developed and validated at the global scale and would require further methodological and data advances. Delineation of aquifer structure and confining units ultimately relies on detailed three-dimensional subsurface information derived from dense lithologic logs and well secondary data, for example, as used by GebreEgziabher et al. (2025) to estimate theoretically recoverable groundwater across the United States.

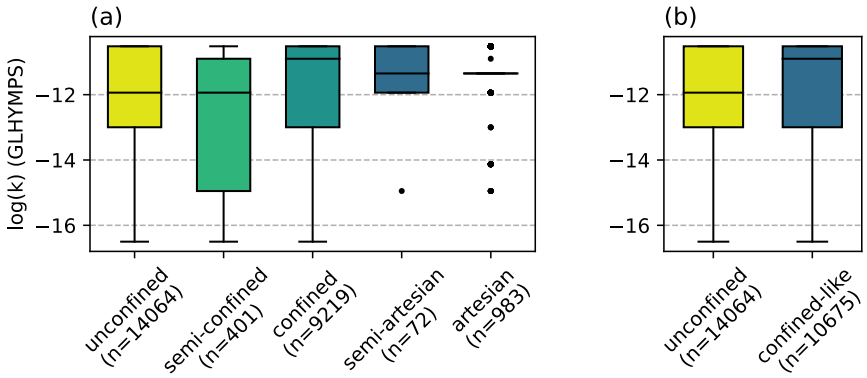


Figure A2: Evaluation of a global-scale near-surface proxy for aquifer state (CGWL subset). (a) GLHYMPS (Huscroft et al. 2018) near-surface permeability distributions for the subset of CGWL wells with reported aquifer state (7%), showing substantial overlap between aquifer-state classes. (b) Same as (a), but grouping semi-confined, confined, semi-artesian, and artesian wells as confined-like.

These constraints in differentiating aquifers at large scales are especially important for interpreting wells that have negative GWLs in the CGWL dataset. In the dataset statistics (Section A.4), about 8,000 wells (~1.3% of wells) have negative median WTDs, many of

which are flagged as unconfined aquifers. It is, however, most often unclear whether these observations reflect artesian heads in confined (or leaky) systems or artefacts introduced by reference-point conversion (Section A.1). Therefore, these negative values are retained in the CGWL dataset.

When using the CGWL dataset to characterize shallow, near-surface groundwater conditions, such negative WTDs can shift summary statistics toward shallower depths and increase estimated groundwater elevations, regardless of whether they represent true above-ground heads or (small) conversion errors. Similarly, surface-based spatial aggregation cannot resolve the three-dimensional structure of groundwater systems or delineate the groundwater contributing area when relating groundwater observations to surface properties. If aggregation combines observations from different aquifer storeys, unit-level summaries of groundwater depth and elevation may reflect a mixture of shallow and deeper conditions rather than the intended near-surface signal. Moreover, the surface aggregation unit may not coincide with the actual area influencing the monitored GWLs.

A.3 Data aggregation

GWL observations were aggregated using the H3 hexagonal spatial indexing system at resolutions 6 and 8 (H3-6 and H3-8; Uber Technologies Inc 2024), which generates near-equal-area grid cells globally with average cell areas of approximately 36 km² and 0.74 km², respectively. These regular grids provide area-representative summaries of groundwater conditions that are straightforward to interpret and reuse in large-scale analyses. Similar gridded products have been developed elsewhere, for example, the 1-km monthly GWL grid for China by Wang et al. (2025). In addition, GWL observations were aggregated within coastal HydroBASINS catchments at level 12 (HYBAS-12; Lehner and Grill 2013) which provide hydrologically defined spatial units widely used in large-scale hydro(geo)logical studies (e.g., Nearing et al. 2024; Kretschmer et al. 2025b).

For spatial aggregation, the preprocessed GWL data (Section A.1) were used. At each well, GWL time series were first aggregated to annual values, and the median of these annual values was then computed for each well (Table A2 in Section A.4). Spatial aggregation followed a two-step procedure. First, wells were summarized to unique location points defined by rounded coordinates to three decimal places, computing location-level median groundwater depth and groundwater elevation (~480,000 unique locations). Second, these location-level values were aggregated to grid cells and basins by computing summary statistics (median, mean, minimum, and maximum) across all locations within each

spatial unit. Aggregation to H3-8, H3-6, and HYBAS-12 units resulted in 232,288, 36,735, and 15,152 grid cells or basins, respectively.

Aggregation can reduce spatial sampling bias, but larger units increase the risk of combining wells that represent different aquifer conditions when aggregation units do not align with hydrogeologic boundaries. Global-scale analyses based on finer spatial units, such as H3-8 grids, remain more strongly influenced by regions with dense monitoring, whereas coarser units distribute observations more evenly but also combine more heterogeneous aquifer conditions. Likewise, the surface-hydrology HYBAS-12 basins do not systematically offer a hydrogeologic advantage, because basin boundaries can still aggregate across poorly resolved aquifer geometries and flow systems, and vertically stacked aquifers (Section A.2). This trade-off is reflected in Figure A3a: the aggregation error, expressed as the RMSE between well-level and unit-level median groundwater depth, is lowest for the H3-8 grids and higher for H3-6 and HYBAS-12 units, and it primarily decreases with increasing numbers of wells per unit. HYBAS-12 basins are on average about four times larger than H3-6 grids (141 km² vs. 36 km²), yet the share of units containing multiple wells is similar for these two aggregation schemes (Figure A3b), and aggregation error is slightly larger for HYBAS-12 basins than for H3-6 grids (Figure A3a).

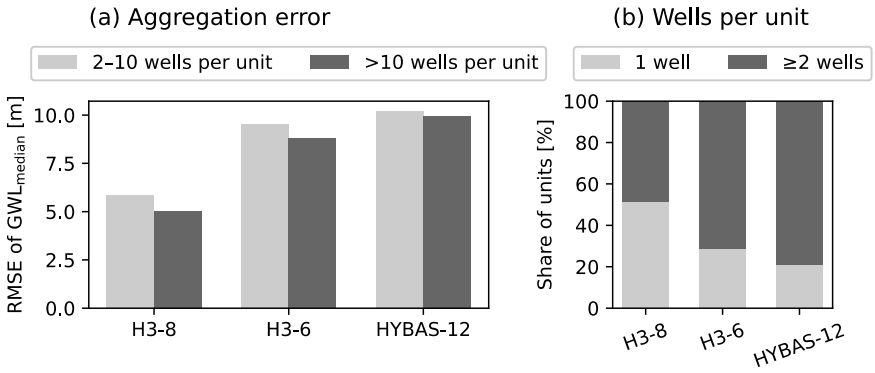


Figure A3: (a) Within-unit RMSE of well-level median groundwater depth for H3-8 and H3-6 grids and HYBAS-12 basins, separated into units with 2–10 and >10 wells. Smaller aggregation units show lower within-unit spread than larger units. (b) Share of units containing exactly one well versus ≥2 wells, illustrating that about half of H3-8 cells are supported by a single well, while the share of units with multiple wells is almost similarly high for H3-6 grids and HYBAS-12 basins.

A.4 CGWL dataset structure

Figure A4 shows the structure of the CGWL dataset. The preprocessed GWL data were aggregated into standardized time intervals (weekly, monthly, annual) and stored in *gwl-*

ts_weekly.csv, *gwl-ts_monthly.csv*, and *gwl-ts_annual.csv* (Table A1). No re-aggregation of already temporally aggregated time series to finer resolutions was performed. Specifically, some wells from source numbers 8 and 14 (Table 1) are only available as annual or monthly aggregates. Summary statistics for each time series, including observation counts, time coverage, and basic metrics, are recorded in the corresponding statistics tables (*statistics_weekly.csv*, *statistics_monthly.csv*, *statistics_annual.csv*; Table A2). Well locations and attributes, including coordinates, elevation data, aquifer descriptors, and well facility information, are compiled in *wells.csv* (Table A3). Metadata on data source references, access dates, and source-specific licensing information are listed in *references.csv* (Table A4). Spatially aggregated groundwater depth statistics are provided in *grid6_groundwater_depth.csv*, *grid8_groundwater_depth.csv*, and *hybas_groundwater_depth.csv* (Tables A7 and A8), while corresponding groundwater elevation aggregates are provided in *grid6_gw_elevation.csv*, *grid8_gw_elevation.csv*, and *hybas_groundwater_elevation.csv* (Tables A9 and A10). The geometries of the aggregation units are stored in *grids_h3_6.gpkg*, *grids_h3_8.gpkg*, and *HYBAS.gpkg* (Tables A5 and A6).

Each time series is uniquely identified by a "well_id", and can be linked to its spatial aggregation unit via the H3 index fields or the basin identifier ("h3_6_index", "h3_8_index", and "HYBAS_ID"). All identifiers are merged at the well level in *wells.csv* (Table A3). The "reference_id" field links each well and its groundwater data to the corresponding source listed in *references.csv* (Table A4). Aggregation, structuring, and calculation of derived fields were conducted using Python (Python Software Foundation 2021) and ArcGIS Pro (Esri 2022).

Table A1: Metadata for the GWL time series (*gwl-ts_weekly.csv*; *gwl-ts_monthly.csv*; *gwl-ts_annual.csv*).

Variable	**Description**	**Type**	**Unit / Format**	**Example**
well_id	Unique identifier for each well consisting of the reference_id followed by a sequential number.	String	Unique identifier	USAAA101579
date	Aggregated date of observation.	Datetime	Datetime [YYYY-MM-DD]	2002-08-26
Value	Groundwater level measured relative to below the ground surface reference point (mbrp). The ground surface reference is either natural ground (preferred) or top of casing.	Float	[m] (2 decimals)	12.27

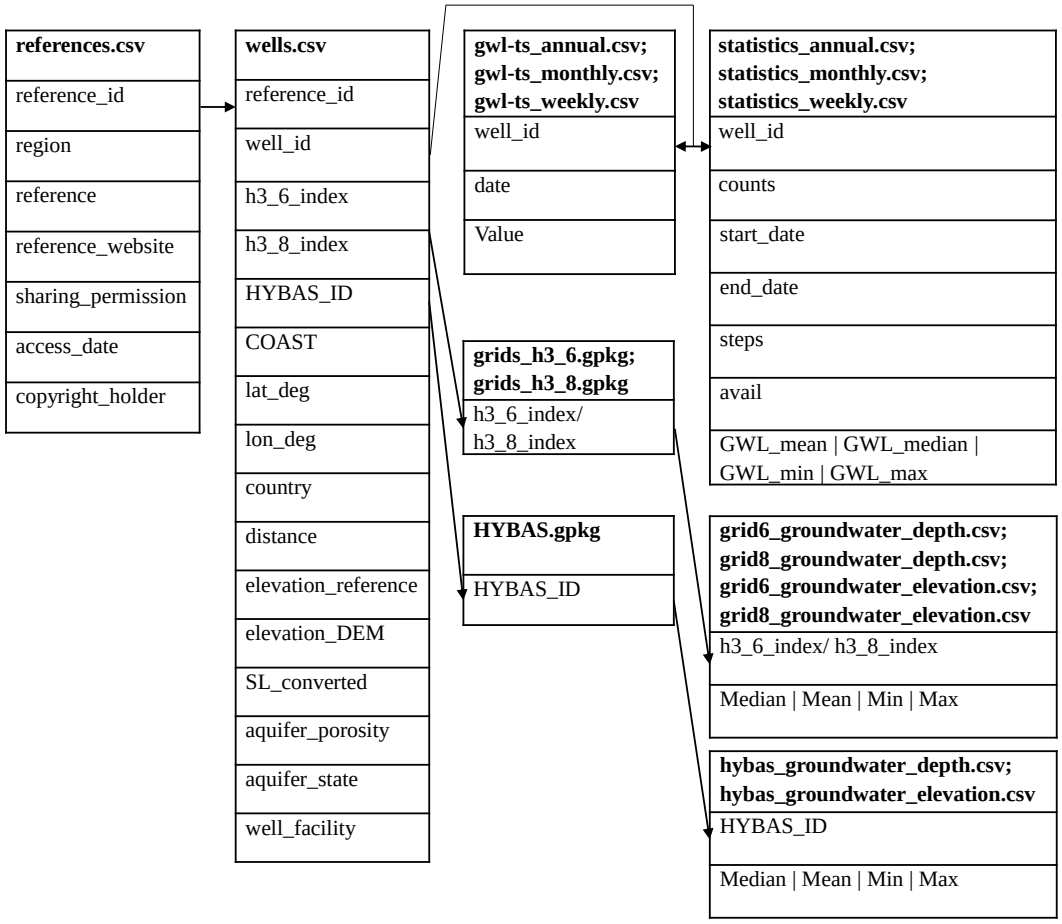


Figure A4: Structure of the CGWL dataset.

Table A2: Metadata for the GWL time series statistics (*statistics_weekly.csv*; *statistics_monthly.csv*; *statistics_annual.csv*).

Variable	Description	Type	Unit / Format	Example
well_id	Unique identifier for each well consisting of the reference_id followed by a sequential number.	String	Unique identifier	USAAA101579
counts	Count of observations in the time series.	Integer	Count	250
start_date	First date of the time series.	Datetime	Datetime [YYYY-MM-DD]	2001-01-01
end_date	Last date of the time series.	Datetime	Datetime [YYYY-MM-DD]	2005-12-29
steps	Number of time steps in the time series.	Integer	Count	261
avail	Availability of data in decimal form, representing the fraction of available time steps.	Float	[-] (2 decimals)	0.95
GWL_mean	Mean value of the groundwater level time series.	Float	[m] (2 decimals)	12.34
GWL_median	Median value of the groundwater level time series.	Float	[m] (2 decimals)	12.30
GWL_min	Smallest groundwater level value in the time series.	Float	[m] (2 decimals)	10.85
GWL_max	Largest groundwater level value in the time series.	Float	[m] (2 decimals)	14.02

Table A3: Metadata for the well locations (*wells.csv*).

Variable	Description	Type	Unit / Format	Example
reference_id	Unique identifier for the reference. Consists of the 3166-1 alpha-2 country code followed by a group identifier for the reference.	String	Unique identifier	USAAA
well_id	Unique identifier for each well consisting of the reference_id followed by a sequential number.	String	Unique identifier	USAAA101579
h3_6_index	Unique identifier for each H3 grid cell at resolution 6, representing a geographic area in the H3 hexagonal grid system[1].	String	H3 index	862a1341fffffff
h3_8_index	Unique identifier for each H3 grid cell at resolution 8, representing a geographic area in the H3 hexagonal grid system[1].	String	H3 index	882a134e6dfffff
HYBAS_ID	Unique identifier for each HydroBASIN at resolution 12[2].	String	HydroBASIN identifier	5120778930
COAST	Flag (1) indicates a coastal HydroBASIN[2].	Integer	[-]	1
lat_deg	Latitude coordinate of the well location.	Float	Degrees [°] (3 decimals)	39.983
lon_deg	Longitude coordinate of the well location.	Float	Degrees [°] (3 decimals)	-75.050
country	3166-1 alpha-2 country code.	String	Code	US
distance	Distance to the nearest coastal segment[3].	Float	Meters [m] (2 decimals)	43.20
elevation_reference	Elevation of the well location as provided by the reference data source. Used to convert between groundwater depth and groundwater elevation.	Float	Meters above sea level [masl] (2 decimals)	7.78
elevation_DEM	Elevation of the well location based on the Digital Elevation Models (DEM) CoastalDEM[4] and NASADEM[5]. Used to convert between groundwater depth and groundwater elevation when elevation_reference was not available.	Float	Meters above sea level [masl] (2 decimals)	8.10
SL_converted	DEM-derived groundwater depth cases were flagged with 1.	Integer	[-]	1
aquifer_porosity	Categorized porosity type of the aquifer, derived from keywords in the raw data. Possible values: pore, karst, fissure.	String	Text descriptor	pore
aquifer_state	Categorized hydrological state of the aquifer, derived from keywords in the raw data. Possible values: unconfined, semi-confined, confined, semi-artesian, artesian.	String	Text descriptor	confined
well_facility	Flag (1) indicating possible human influence, such as pumping or artificial recharge, based on keywords in the raw data.	Integer	[-]	1

Table A4: Metadata for the references (*references.csv*).

Variable	Description	Format
reference_id	Unique identifier for the reference. Consists of the 3166-1 alpha-2 country code followed by a group identifier for the reference.	String
region	Name of the geographical region or country of the reference.	String
reference	Name of the data provider, given in the form: A – B (C), where A is the portal/application (if applicable), B is the responsible authority, and C is the jurisdiction.	String
reference_website	Website for data access or contact information.	String
sharing_permission	Details about the sharing or licensing permissions applicable to the data.	String
access_date	Date when the data was accessed.	Datetime [YYYY-MM-DD]
copyright_holder	Multiple copyright holders where explicitly stated in the source licence/terms, otherwise responsible authority (reference).	String

Table A5: Metadata for the grids (*grids_h3_6.gpkg*; *grids_h3_8.gpkg*).

Variable	Description	Type	Unit / Format	Example
h3_6_index / h3_8_index	Unique identifier for each H3 grid cell at resolution 6 / 8, representing a geographic area in the H3 hexagonal grid system[1].	String	H3 index	862a1341fffffff / 882a134e6dfffff
geometry	Geospatial representation of the hexagonal grid's location, stored as a polygon in the GPKG file.	Geometry	Polygon [GPKG]	Polygon

Table A6: Metadata for the basins (*HYBAS.gpkg*).

Variable	Description	Type	Unit / Format	Example
HYBAS_ID	Unique identifier for each HydroBASIN at resolution 12[2].	String	HydroBASIN identifier	5120778930
geometry	Geospatial representation of the basin, stored as a polygon in the GPKG file.	Geometry	Polygon [GPKG]	Polygon

Table A7: Metadata for the groundwater depth in grids (*grid6_groundwater_depth.csv*; *grid8_groundwater_depth.csv*).

Variable	Description	Type	Unit / Format	Example
h3_6_index / h3_8_index	Unique identifier for each H3 grid cell at resolution 6 / 8, representing a geographic area in the H3 hexagonal grid system[1].	String	H3 index	862a1341fffffff / 882a134e6dfffff
Median	Median groundwater depth for each H3 grid cell, derived from well data within the grid.	Float	Meters [m] (2 decimals)	3.45
Mean	Mean groundwater depth for each H3 grid cell, derived from well data within the grid.	Float	Meters [m] (2 decimals)	3.40
Min	Minimum groundwater depth for each H3 grid cell, derived from well data within the grid.	Float	Meters [m] (2 decimals)	3.33
Max	Maximum groundwater depth for each H3 grid cell, derived from well data within the grid.	Float	Meters [m] (2 decimals)	3.46

Table A8: Metadata for the groundwater depth in basins (*hybas_groundwater_depth.csv*).

Variable	Description	Type	Unit / Format	Example
HYBAS_ID	Unique identifier for each HydroBASIN at resolution 12[2].	String	HydroBASIN identifier	5120778930
Median	Median groundwater depth for each basin, derived from well data within the basin.	Float	Meters [m] (2 decimals)	3.45
Mean	Mean groundwater depth for each basin, derived from well data within the basin.	Float	Meters [m] (2 decimals)	3.40
Min	Minimum groundwater depth for each basin, derived from well data within the basin.	Float	Meters [m] (2 decimals)	3.33
Max	Maximum groundwater depth for each basin, derived from well data within the basin.	Float	Meters [m] (2 decimals)	3.46

Table A9: Metadata for the groundwater elevation in grids (*grid6_groundwater_elevation.csv*; *grid8_groundwater_elevation.csv*).

Variable	Description	Type	Unit / Format	Example
h3_6_index / h3_8_index	Unique identifier for each H3 grid cell at resolution 6 / 8, representing a geographic area in the H3 hexagonal grid system[1].	String	H3 index	862a1341fffffff / 882a134e6dfffff
Median	Median groundwater elevation for each H3 grid cell, derived from well data within the grid.	Float	Meters [m] (2 decimals)	3.45
Mean	Mean groundwater elevation for each H3 grid cell, derived from well data within the grid.	Float	Meters [m] (2 decimals)	3.40
Min	Minimum groundwater elevation for each H3 grid cell, derived from well data within the grid.	Float	Meters [m] (2 decimals)	3.33
Max	Maximum groundwater elevation for each H3 grid cell, derived from well data within the grid.	Float	Meters [m] (2 decimals)	3.46

Table A10: Metadata for the groundwater elevation in basins (*hybas_groundwater_elevation.csv*).

Variable	Description	Type	Unit / Format	Example
HYBAS_ID	Unique identifier for each HydroBASIN at resolution 12[2].	String	HydroBASIN identifier	5120778930
Median	Median groundwater elevation for each basin, derived from well data within the basin.	Float	Meters [m] (2 decimals)	3.45
Mean	Mean groundwater elevation for each basin, derived from well data within the basin.	Float	Meters [m] (2 decimals)	3.40
Min	Minimum groundwater elevation for each basin, derived from well data within the basin.	Float	Meters [m] (2 decimals)	3.33
Max	Maximum groundwater elevation for each basin, derived from well data within the basin.	Float	Meters [m] (2 decimals)	3.46

[1] Uber Technologies Inc (2024)

[2] Lehner and Grill (2013)

[3] Sayre et al. (2021)

[4] Kulp and Strauss (2019)

[5] NASA JPL (2020)

A.5 Spatiotemporal coverage

The CGWL dataset provides extensive spatial coverage of coastal regions worldwide. However, observation density is strongly concentrated in the Northern Hemisphere (Figure A5), with particularly high data availability in North America, northwestern Europe, and Australia. In contrast, several regions remain underrepresented or entirely uncovered, such as most of Africa, the equatorial regions, large parts of South America, Asia, and many small islands. Records in the CGWL dataset extend back to the 1800s but remain sparse before the 1970s (Figure A6). There is a noticeable increase in observations from the 1990s onward, peaking between 2008 and 2012, before gradually declining toward the end of the data collection period of the CGWL dataset (between 2019 and 2025). The dataset also shows a distinct seasonal pattern, with consistently higher counts in some months, such as March and October. However, because the data originate from diverse regions worldwide, these patterns do not reflect a specific seasonal focus of the CGWL dataset as a whole. To assess time-series completeness across temporal aggregations, Figure A7 summarizes key statistics for the weekly, monthly, and annual products: the number of available observations (counts), the number of expected time steps given the aggregation frequency (steps), and data availability (avail). Weekly aggregations have substantially more expected time steps, but lower data availability, whereas monthly series show intermediate completeness and annual series are typically the most complete and regular (Figure A7).

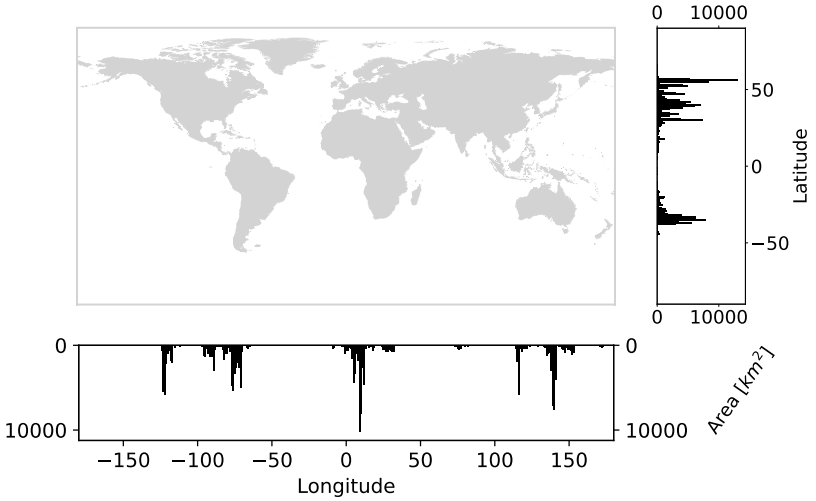


Figure A5: Spatial coverage of CGWL observations represented by H3 grid cells (resolution 8). The marginal bar plots show the cumulative surface area of data-covered grid cells aggregated by latitude (right) and longitude (bottom), reported in km².

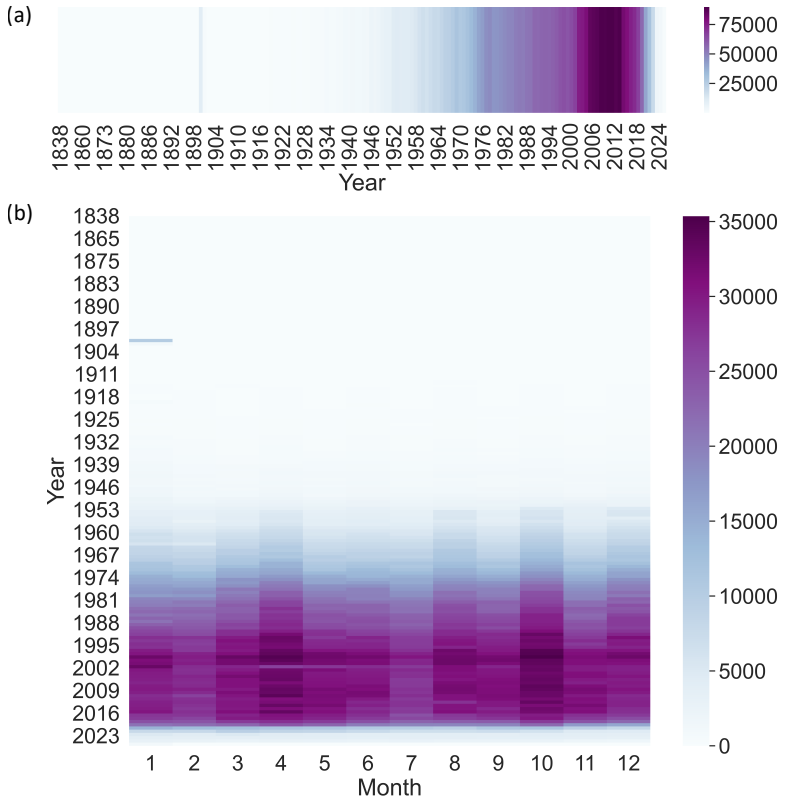


Figure A6: Temporal distribution of GWL observations in the CGWL dataset. (a) Annual data availability by year. (b) Monthly data availability by year and month. Each vertical stripe represents the number of wells with observations in a given year (and month).

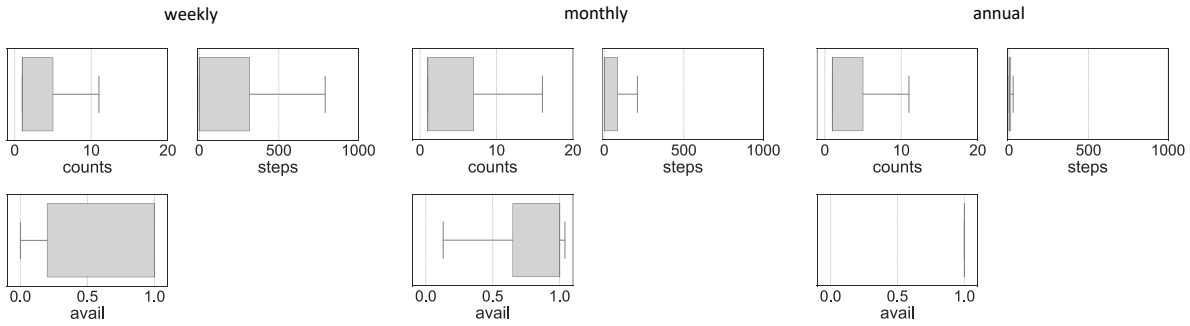


Figure A7: Boxplots of temporal statistics for GWL time series in the CGWL dataset at three temporal aggregation levels: weekly, monthly, and annual. The plots show the distribution of counts (number of available observations), steps (number of expected time steps based on the aggregation frequency and the start and end dates of each time series), and avail (data availability, expressed as the fraction of time steps with observations). Outliers are excluded for visual clarity.

A.6 Source-specific details on data acquisition and preprocessing

GWL data for the CGWL dataset were acquired from the sources listed in Table 1. Table A11 provides source-specific notes that summarize how data were acquired and preprocessed for each region, reflecting experience gained during compilation of the CGWL dataset. These notes also include information on the availability of secondary information (Section A.1). Where available, such secondary data were reviewed and considered for integration into the CGWL dataset (*wells.csv*; Table A3). For some sources, additional contextual materials (e.g., borehole logs or drilling reports) are available, but they are heterogeneous, not consistently digitized, and typically require case-specific interpretation. Therefore, they were not systematically integrated.

Table A11: Overview of source-specific data acquisition and preprocessing decisions in compiling the CGWL dataset, including available secondary attributes (e.g., reference elevations, well purpose, aquifer type). Source numbers and regions correspond to those listed in Table 1.

Number(s) / region	Description
1 Australia	GWL data were downloaded on 11/4/2021 from the Australian Groundwater Explorer (http://www.bom.gov.au/water/groundwater/explorer) hosted by the Australian Government – Bureau of Meteorology. The downloaded version was "data VERSION 1.7.0" and included GWL measurements for individual states (data owners), along with secondary information from the National Groundwater Information System (NGIS). The reference point for GWL measurements varied by state and included ground surface, top of casing, and sea level. Secondary attributes were extracted from the geodatabase into CSV files and linked to the GWL data using the "BoreID" field, enabling the association of well locations and secondary information. Coordinates were originally provided in the Geocentric Datum of Australia 1994 (GDA94). The fields "FType" and "FTypeClass" were used to extract the purpose of each well. Aquifer information was not available at the individual well level, but could in principle be inferred via the hydrogeologic unit associated with each bore. However, units were considered too coarse for the well-level detail required in the CGWL dataset.
2 Belgium	GWL data from Flanders in Belgium were downloaded on 9/23/2020 from Databank Ondergrond Vlaanderen (DOV; https://www.dov.vlaanderen.be) using the Python package "pydov". Coordinates were originally provided in the Belgian Lambert 72 system. GWL measurements were referenced to sea level, but ground surface elevation was available for conversion. Information on aquifer confinement and the purpose of the well filter were derived from the "Regime" and "filtertype" fields, respectively.
3 Brazil	GWL data from Brazil were downloaded on 9/15/2022 through the online data portal RIMAS (http:/rimasweb.cprm.gov.br/layout/pesquisa_complexa.php?rimas=true, operated by the Serviço Geológico do Brasil, using Python. The dataset did not report well purpose explicitly, but it consisted of wells from the national monitoring network. Information about aquifer type was derived from the field "Aquífero".

Continued on next page

continued.

Number(s) / region	Description
4–6 Canada	GWL data from Canada were obtained from three sources: the Nova Scotia and British Columbia provincial government agencies (https://www.novascotia.ca/; https://www2.gov.bc.ca/gov/), and the Groundwater Information Network (GIN; http://gw-info.net). GWL data from several provinces were first obtained via a download file from GIN on 3/30/2020. In this dataset, no reference point was specified for GWL observations from British Columbia. In the Nova Scotia data, the date format suggested that day and month fields were swapped. To address these uncertainties, GWL data referenced to the ground surface were downloaded from British Columbia (http://www.env.gov.bc.ca/wsd/data_searches/obswell/map/data/ObservationWellDataAll_DailyMean.csv) on 3/1/2021 and merged with well locations from the GIN dataset using original identifiers. On 8/4/2021, GWL data and well locations were downloaded from the Nova Scotia website (https://novascotia.ca/nse/groundwater/groundwaternetwork.asp). In this dataset, GWLs were referenced to sea level, and an elevation offset was available to convert values to the top of casing. A small number of GWL records from Quebec and Prince Edward Island were retained from the GIN dataset for coastal regions, with GWLs referenced to sea level. The GIN dataset did not contain information on aquifer type or confinement state. The Nova Scotia dataset included aquifer type but not confinement state for most wells. No information was available regarding the well's purpose. However, all data originated from monitoring networks.
7 Denmark	GWL and well data were extracted from the Jupiter database of the Geological Survey of Denmark and Greenland (GEUS; https://www2.gov.bc.ca/gov/) on 5/16/2019 (https://www.geus.dk/produkter-ydelser-og-faciliteter/data-og-kort/national-boringsdatabase-jupiter/). The tables "watlevel", "borehole" and "intake" were accessed. GWLs provided with the "watlevel" table were merged with "borehole" locations via the "intake" table. The reference point for the GWL data was different throughout the datasets, including ground surface, top of casing, and sea level. Information about the bore purpose was available in the field "PURPOSE". Information on the aquifer confinement state was available in the field "WATERTABLETYPE".
8 Europe	GWL data were obtained from the Waterbase – Water Quantity dataset (https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quantity-14), which contains time series reported by member and cooperating countries of the European Environment Agency (EEA). The dataset was downloaded on 7/13/2021. GWL is reported "as the mean value during the reported period (one month, one quarter, or one year)", and "the reference for groundwater level value is the mean sea level" (https://dd.eionet.europa.eu/dataelements/78643). Only GWL time series with a monthly resolution were selected. Data from Greece and Cyprus showed suspect GWL values, likely due to a decimal error, and were removed, along with data from three individual stations with unrealistic values. In total, GWL data from the following countries were used: Belgium, Bulgaria, Denmark, Estonia, Ireland, Italy, Latvia, Lithuania, Slovenia, Spain, and Sweden. The purpose of Waterbase is to report on the status and quality of water resources. The dataset does not contain information on aquifer type or confinement state.
9 France	GWL data and secondary attributes from France were obtained from the ADES database (https://ades.eaufrance.fr/Recherche) on 10/30/2020. GWLs were referenced to ground surface, top of casing, and sea level. Where well elevation was available, GWLs were converted from sea level to ground surface reference. Ground surface elevation was mainly available. Information on the aquifer type and confinement state was obtained from the fields "Mode de gisement" and "Caractéristiques aquifère", respectively. Information on the well purpose was extracted from the field "Usage".

Continued on next page

continued.

Number(s) / region	Description
10–13 Germany	GWL data and secondary attributes were obtained from four government agencies of the German coastal regions between 11/24/2020 and 4/19/2021: (a) Landesamt für Umwelt (previously: Landesamt für Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein (LLUR); https://www.schleswig-holstein.de/DE/landesregierung/ministerien-behoerden/LLUR/llur_node.html), (b) Niedersächsischer Landesbetrieb für Wasserwirtschaft, Küsten- und Naturschutz (NLWKN; https://www.nlwkn.niedersachsen.de/startseite/), (c) Staatliches Amt für Landwirtschaft und Umwelt Mittleres Mecklenburg (https://www.stalu-mv.de/mm), and (d) Staatliches Amt für Landwirtschaft und Umwelt Vorpommern (https://www.stalu-mv.de/vp). GWLs were referenced either to sea level or to ground surface, and ground surface elevation was partly available for conversion. Well coordinates were provided in projected coordinate systems for Germany. Information on aquifer type (but not confinement state) was available from source number 11 (Table 1). All datasets originate from governmental monitoring networks.
14 Global	GWL data were obtained from two IGRAC-hosted global datasets on 1/26/2025. First, data from the Global Groundwater Monitoring Network (GGMN) were downloaded through IGRAC (IGRAC 2025). This dataset compiles GWL records from multiple national and regional agencies worldwide. Although numerous secondary fields were included, most lacked data entries. The GGMN data considered for integration into the CGWL dataset were compiled from: Environment and Energy Administration of the Brussels-Capital Region, Belgium; Geological Survey of Brazil; Geological Survey Canada; National Geological Surveys for Denmark and Greenland (GEUS); Ministry for the Ecological Transition and the Demographic Challenge, Spain (MITECO); Geological Survey of Estonia; French Geological Survey (BRGM); British Geological Survey (BGS); Department of Water Resources, Gambia; Central Ground Water Board, India; Environmental Protection Agency, Ireland; Hydrological Service, Israel; "Ministry of Agriculture, Water and Forestry, Department of Water Affairs and Forestry, Division Water Environment", Namibia; Norwegian Water Resources and Energy Directorate (NVE); National Water Resources Information System, Portugal (SNIRH); Ministry of Water Resources, Sierra Leone; Ministry of Environment, El Salvador; Geological Survey of Sweden; Water Resources Agency Taiwan (MOEA); Western Cape University (South Africa). Second, annual GWL data associated with Jasechko et al. (2024) were downloaded from the IGRAC dataset page.
15 Spain	GWL data from Spain were obtained for the Ebro River region from the Instituto Geológico y Minero de España (https://sig.mapama.gob.es/redes-seguimiento/?herramienta=Sondeos) on 3/17/2020, using Python and well identifiers available on the website of the regional governmental agency (http://iber.chebro.es/geoportal/). Well coordinates were originally reported in a projected coordinate system for Spain. GWL data referenced to the ground surface were used. No information on aquifer type or confinement state was available in the dataset (only the aquifer name was provided). Some specifications regarding the well purpose were available in the field "Uso".
16 India	GWL time series and well data from India were downloaded from India Water Resource Information System (India-WRIS; http://indiawris.gov.in/wris/#/DataDownload) on 2/26/2020. The fields "DISTRICT", "SITE_NAME", and "STATION" were used to merge the GWL and well datasets. Email correspondence confirmed that the vertical reference point for GWL measurements is the top of casing. No secondary information was available in the dataset. The data originate from monitoring wells with measurements recorded four times per year.

Continued on next page

continued.

Number(s) / region	Description
17 Ireland	GWL data from Ireland were downloaded from the Environmental Protection Agency (EPA; https://epawebapp.epa.ie/hydronet/#Groundwater) on 2/26/2020 using Python. Well location data were provided by the EPA via email. Derived GWLs (except for "DIP" wells) were referenced to sea level, and no well elevation was available for conversion at the time the dataset was accessed. Well coordinates were available in a projected coordinate system for Ireland. Information on aquifer type was manually extracted from the EPA website. The dataset consists of monitoring wells from the national network, however, potential impacts from nearby pumping wells are indicated on the website.
18–19 South Africa	GWL data were obtained from two sources: (a) data from the National Groundwater Archive (NGA; https://www.dws.gov.za/NGANet/Security/WebLoginForm.aspx) were received on 6/12/2019 following an email request to NGAUR@dws.gov.za, and (b) GWL time series from the national Hydstra database were obtained between 3/27/2021 and 4/1/2021 via email request to HydstraS@dwa.gov.za. Email correspondence confirmed that the reference point for GWL observations in South Africa is the top of casing. No information on aquifer type or confinement state was available in the datasets. GWL data from Hydstra are part of the national monitoring network. For the NGA dataset, well purpose may differ from monitoring, as indicated by the field "GeositeInfo_GeositeType".
20 Sweden	GWL time series from Sweden were obtained on 10/22/2021 from the Geological Survey of Sweden (SGU) using Python, following the instructions provided at https://www.sgu.se/produkter/geologiska-data/oppna-data/grundvatten-oppna-data/grundvattennivaer-tidsserier/. GWLs were reported with reference to either ground surface or top of casing. Well coordinates were given in a projected coordinate system for Sweden. Information on aquifer confinement state was available in the field "Akvifertyp". According to email correspondence with SGU (KundserviceSGU@mail.sgu.se; as of 10/20/2021), all the data comes from observation sites that have been drilled for observation purposes.
21 The Netherlands	GWL time series and well data from the Netherlands were obtained from the DINOloket database (https://www.dinoloket.nl) on 8/25/2020. Well coordinates were reported in a projected coordinate system for the Netherlands. The GWL data referenced to ground surface, available for all wells, were used. No secondary information was available in the dataset.
22–23 United States	GWL time series and well data from the United States were obtained from two sources: (a) the United States Geological Survey (USGS; https://waterdata.usgs.gov/nwis/gw; https://waterservices.usgs.gov/rest/GW-Levels-Test-Tool.html) on 5/26/2020, and (b) the California Department of Water Resources (https://wdl.water.ca.gov) on 6/16/2022. GWLs from the USGS were reported with reference to either ground surface or sea level. Well coordinates were reported in the NAD83 geodetic coordinate system for North America in both datasets. Information on aquifer confinement state was available in the USGS dataset, but no information on aquifer type or well purpose. The dataset from California did not include information on aquifer type or confinement state, but included well purpose in the field "WELL_USE".

Appendix B Coastal groundwater level trends reveal global susceptibility to seawater intrusion

B.1 Time series selection and filtering criteria for trend analysis

To analyze recent GWL trends at the global scale, we filtered monthly CGWL time series using 9- and 19-year windows within the period 1990–2024. For each well, the longest valid segment was selected. If a single segment satisfied both durations, the 9-year trend is nested within the final portion of the 19-year segment. Otherwise, only the 9-year trend was computed. These time series lengths represent a pragmatic compromise between the record lengths ideally needed for robust long-term trend assessment (often $\geq$ 30 years; Section B.3) and the limited availability of sufficiently long, continuous GWL observations across regions worldwide (see Appendix A). In addition, we applied data-availability rules in the time-series selection process that allow for variable observation patterns while limiting seasonal distortion: (1) no more than 9 months between at least two consecutive observations, and (2) allowance for irregular gaps up to 36 months. An exception was made for GWL data from the Jasechko et al. (2024) dataset included in source 14 in Table 1. These data were available at annual resolution and were therefore processed separately using: (1) no more than 1 year between at least two consecutive observations, (2) allowance for irregular gaps up to 2 years, and (3) a minimum of 5 annual observations. In total, 90,967 time series were retained for the 9-year trends and 32,450 for the 19-year trends. These correspond to 51,263 and 20,857 unique locations, defined as distinct combinations of latitude and longitude at three decimal places. Figure B1 summarizes the temporal distribution of all GWL time series (including duplicated locations) for the 9-year and 19-year windows. For both trend windows, data coverage is highest between 2004 and 2011, and data density declines toward both edges of the selection period.

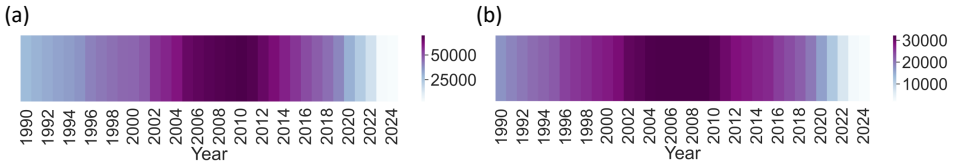


Figure B1: Temporal distribution of available years in GWL time series analyzed for trends across all wells selected for the 9-year (a) and 19-year (b) windows. Each vertical stripe represents the number of wells with valid observations in a given year.

B.2 Spatial availability of trend results

After spatial aggregation using the H3 hexagonal spatial indexing system at resolutions 8 Uber Technologies Inc (2024), trend results were available for 36,579 grid cells for the 9-year window and 17,703 grid cells for the 19-year window, meaning that 19-year trends support ~41% of all grid-level trend estimates globally. Their spatial distribution varies strongly across IPCC regions, and the availability of 19-year trends is markedly lower (and in some regions absent) compared to 9-year trends (Figure B2). Most trend results originate from data-rich coastal regions in North America, Europe, India, and Australia, consistent with the spatial sampling patterns of the CGWL dataset (Appendix A). In 13 of 23 IPCC regions, the 19-year analysis has >50% fewer samples than the 9-year analysis or <10 samples over 19 years. To balance temporal robustness and spatial coverage, we prioritized 19-year trends in subsequent analyses and complemented them with 9-year trends where 19-year records are unavailable. In Section B.3, we further discuss the sensitivity of trend estimates to record length and compare 9- and 19-year trends for the overlapping subset of wells to quantify the uncertainty and robustness of using shorter records in global-scale groundwater assessments.

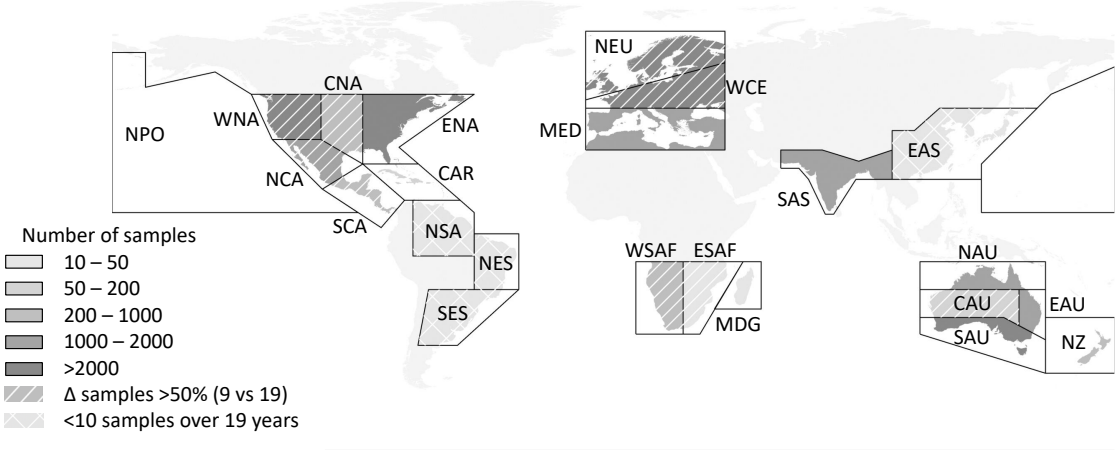


Figure B2: Number of grid cells ("observations") with valid 9-year GWL trends in each IPCC coastal region. Grey shading indicates sample-size classes. Single hatching marks regions with >50% fewer grid cells available for 19-year trends than for 9-year trends, and cross-hatching marks regions with <10 grid cells available over 19 years.

B.3 Sensitivity of trend estimates to record length

Longer time series are generally preferable for robust trend analysis because aquifers tend to attenuate high-frequency variability and preserve low-frequency signals shaped by

climate cycles and aquifer response dynamics (Hannaford et al. 2013; Baulon et al. 2022b; Baulon et al. 2022a). At the same time, the records most widely available at large spatial scales are often short, contain gaps, or have seasonally biased sampling, particularly in global studies such as this one and Jasechko et al. (2024), which used ≥8-year records. Low-frequency components that are often tied to multiannual and decadal variability can, however, strongly influence trend outcomes and common statistical assumptions in trend detection methods (Serinaldi et al. 2018) are often violated in GWL time series. For example, two wells with different 9-year windows may yield different slopes because they sample different phases of macroclimatic oscillations such as El Niño Southern Oscillation (ENSO) that have typical time scales of 2–8 years and can affect GWLs (Abhik et al. 2025; Kuss and Gurdak 2014). As a result, inconsistent trends in the same time series or nearby wells encountered in this and previous studies (Lischeid et al. 2021; Jasechko et al. 2024) are typical phenomena in groundwater research and can arise from timing effects in addition to spatial heterogeneity in aquifer properties and human pressures. Even in data-rich regions, attributing observed trends to specific drivers is difficult because monitoring networks, hydrogeologic information, and records of pumping or recharge interventions are sparse and inconsistent. Consequently, as also emphasized by Jasechko et al. (2024), large-scale syntheses can at best relate GWL trends statistically to covariates such as aridity or cultivation, but cannot robustly disentangle the relative roles of climate variability, groundwater withdrawal, and management interventions for individual aquifer systems. Correlations between trends and these variables should thus not be interpreted as direct causation.

To evaluate how strongly the choice of record length affects our results, we compared trends from overlapping 9- and 19-year windows for the subset of wells that meet both length criteria (Figure B3). At the grid-cell level, 9- and 19-year slopes are moderately correlated (Spearman $\rho_s = 0.58$; Figure B3a). Around three-quarters of grid cells (76%) are classified as consistent, meaning that the 9- and 19-year slopes do not change sign and thus yield the same qualitative interpretation (declining, approximately stable, or rising GWLs). Around one fifth (19%) are classified as ambiguous, where the slopes have opposite signs but at least one has a small magnitude ($|\text{slope}| < 0.1\,\text{m}\,\text{yr}^{-1}$), so that the apparent disagreement is largely driven by near-zero trends. The remaining 5% show clearly inconsistent behaviour, with opposite signs and magnitudes $|\text{slope}| \geq 0.1\,\text{m}\,\text{yr}^{-1}$ in both periods (Figure B3a). Overall, the window-length comparison shows that our global-scale conclusions are robust, while flagging a small subset of locations where the inferred trend direction depends on record length and should therefore be interpreted cautiously.

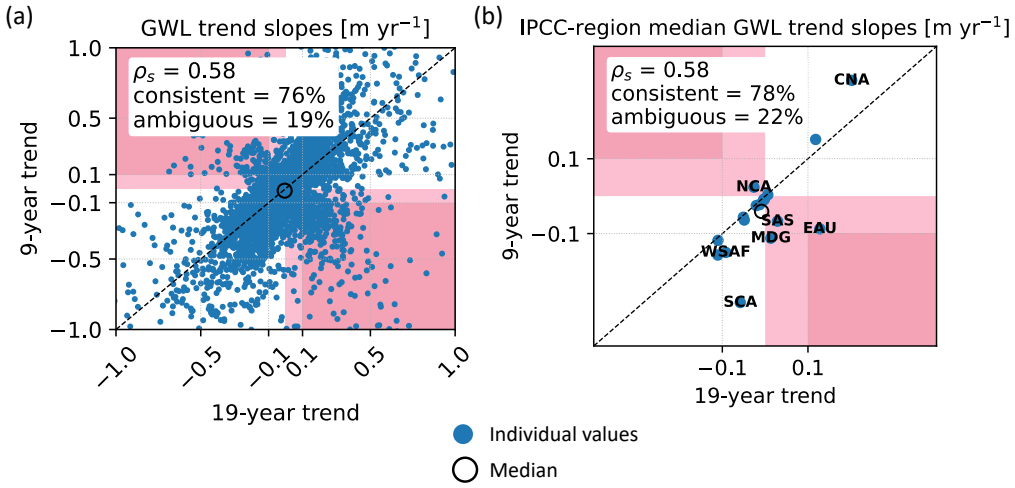


Figure B3: Comparison of 9- and 19-year GWL trend slopes. (a) Grid-cell trend slopes for all cells with both trend windows available. (b) Median grid-cell GWL trend slope within each IPCC region. Labels are shown for regions where $|\Delta\text{slope}| > 0.05\,\text{m}\,\text{yr}^{-1}$. In both panels, red shading marks combinations where 9- and 19-year trend slopes have opposite signs (darker where both $|\text{slope}| \geq 0.1\,\text{m}\,\text{yr}^{-1}$). The Spearman's rank correlation coefficient (ρ_s) and the fractions of consistent trends (no sign reversal between 9- and 19-year slopes) and ambiguous trends (opposite signs but at least one slope of small magnitude, $|\text{slope}| < 0.1\,\text{m}\,\text{yr}^{-1}$) are shown in the inset.

We therefore also examined whether the regional narrative could change with trend-window length. Using the median grid-cell GWL trend slope within each IPCC coastal region as a region-level summary shows overall similar behaviour between the 9- and 19-year windows ($\rho_s = 0.58$; 78% consistent and 22% ambiguous; Figure B3b), but highlights individual regions where trends differ more strongly than elsewhere. Regions that plot in the red-shaded opposite-sign areas in Figure B3b are particularly noteworthy, as they indicate that 9- and 19-year median regional slopes differ in sign. Such sign switches occur in NCA, SAS, MDG, and EAU. The EAU region stands out with negative 9-year median trends but positive 19-year median trends on average. In addition, extreme trends are less frequent over 19-year windows than over 9-year windows (e.g., in SCA, WSAF, and CNA), which is consistent with longer records integrating low-frequency variability and episodic events and therefore dampening short-term extremes. These comparisons indicate that most 9-year records provide a useful first-order signal of longer-term behaviour, but that a non-negligible fraction of wells – and in a few cases whole regions – exhibit trend estimates that are sensitive to the chosen time window. The results underscore the importance of extended GWL records for robust, reliable trend assessments.

For our main analyses, we therefore used a preferred trend diagnostic that combines both

windows, assigning 19-year slopes wherever available and falling back to 9-year slopes otherwise (Section B.2). At the regional scale, mixed-window summaries typically lie between the 9- and 19-year estimates and tend to resemble the 19-year results where a large share of grid cells is supported by 19-year records, while elsewhere, they more closely reflect the 9-year estimates (Tables B1–B3; Section B.5). For example, in EAU the mixed prevalence of significant trends (41%) falls between the 9-year (53%) and 19-year (35%) estimates. In contrast, in the low-sample region MDG, the mixed up/down balance falls outside the 9- to 19-year range (75% upward vs. 55% and 50% in the 9- and 19-year windows; Tables B1–B3), illustrating sensitivity to a few grid cells. Hence, the robustness of regional summaries depends not only on record length but also on regional sample size. We therefore report regional trend results only where $\leq$10 grid cells are available, while noting that substantially larger samples yield more stable region-wise estimates. Among the regions where 9- and 19-year median slopes differ in sign (Figure B3b), only NCA has a majority of grid-cell trends based on 9-year records (Figure B2). Globally, the mixed-window trend product shows 24% significant trends ($|\text{slope}| \geq 0.1\,\text{m}\,\text{yr}^{-1}$), of which 46% are upward and 54% downward. The 9-year window yields 28% significant trends (42% upward, 58% downward), whereas the 19-year window yields 21% significant trends (49% upward, 51% downward).

B.4 Sensitivity of results to spatial aggregation

Building on the discussion in Appendix A of the trade-offs inherent to spatial aggregation, we evaluated how sensitive our diagnostics are to the choice of aggregation unit. We recomputed all analyses using coastal HydroBASINS catchments (HYBAS-12; Lehner and Grill 2013) instead of H3-8 grid cells (Uber Technologies Inc 2024). We compared the results at the level of IPCC regions (Figure B4a–d) and reproduced Figure 6 from the publication (Figure B4e-f). For all four metrics – the share of flat hydraulic gradients, the share of units with trend magnitudes exceeding $\pm 0.1\,\text{m}\,\text{yr}^{-1}$, the fraction of downward trends among trend-active units, and the share of persistent or emerging flat-gradient (not stabilizing) hotspots – values closely follow the 1:1 line, with correlations between $\rho_s = 0.87$–0.95. Globally, switching from grid cells to HYBAS basins changes the results only slightly. Basin aggregation yields a marginally higher prevalence of flat gradients (38% compared to 37%) and classifies fewer units as "no trend" in the 9-year window (64% vs 72%), shifting more units into both upward and downward trend classes (19-year trend-class shares are nearly unchanged). This modest increase in "trend-active" classifications is consistent with the larger spatial support of basins, which aggregates more well-level trend estimates within each unit and, because unit-level trends are summarized as the

median of significant well-level slopes (excluding "no trend" wells), reduces the frequency of units classified as "no trend". Figure B4e-f further indicates that the largest support-related differences in trend activity occur within cluster C3 (flat and energy-limited). Distance-to-coast patterns remain broadly consistent between aggregation schemes.

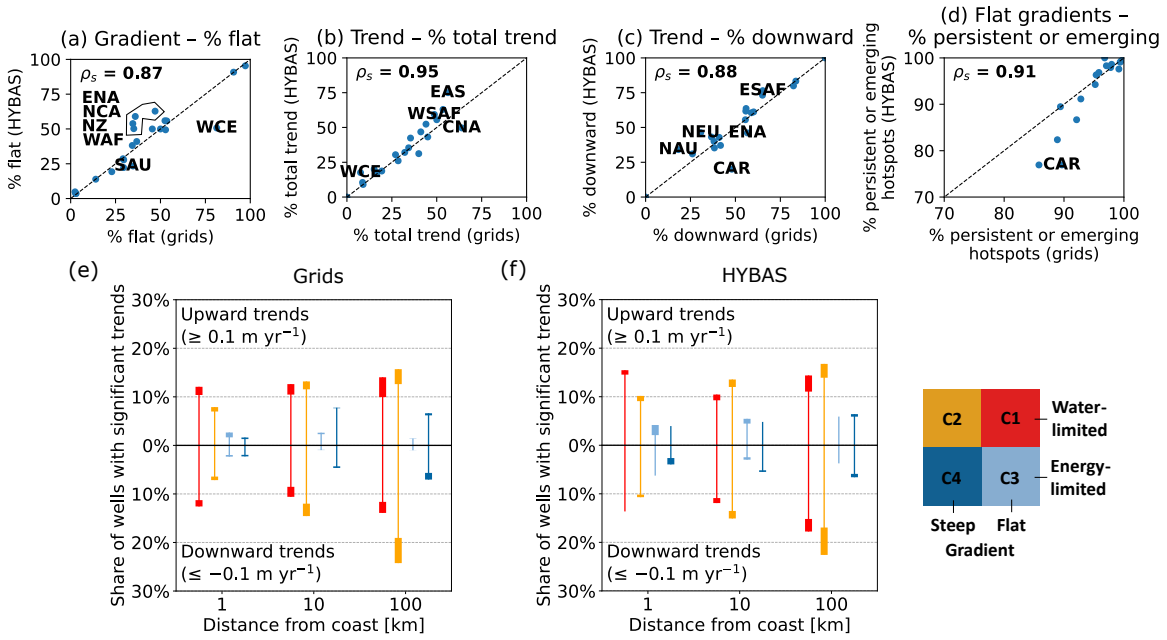


Figure B4: Sensitivity of key diagnostics for SWI susceptibility to the choice of spatial aggregation unit. (a-d) Comparison of H3-8 grid-based and HYBAS-12-based regional statistics for IPCC coastal regions: (a) percentage of units with flat hydraulic gradients, (b) percentage of units with significant trends (absolute slope $\geq 0.1\,\mathrm{m\,yr^{-1}}$), (c) percentage of significant trends that are downward, and (d) percentage of flat-gradient units classified as persistent or emerging hotspots. Each point is one region; the dashed line shows the 1:1 line, and labels mark regions where HYBAS and grid estimates differ by more than 10 percentage points. (e-f) Share of observations with significant upward (top, $\geq 0.1\,\mathrm{m\,yr^{-1}}$) and downward (bottom, $\leq -0.1\,\mathrm{m\,yr^{-1}}$) trends as a function of distance from the coast (1, 10, 100 km), shown for H3-8 grids (e; reproduced from Figure 6) and HYBAS-12 basins (f) to enable direct comparison. Thicker bar segments at the top and bottom indicate the shares of strongly rising ($> 0.5\,\mathrm{m\,yr^{-1}}$) and strongly declining ($< -0.5\,\mathrm{m\,yr^{-1}}$) observations, respectively. Differently colored shares refer to the SWI susceptibility clusters (C1-C4).

Most regions differ by less than 10 percentage points between the two aggregation schemes in the comparison, but 1–6 IPCC regions show larger deviations depending on the analysis. The exceptionally higher share of flat gradients in the grid-based aggregation compared to HYBAS in Western Europe (WCE; Figure B4a) likely reflects the strong influence of densely sampled, low-lying coastal areas (notably the Netherlands) in the grid product. This highlights how differences between aggregation schemes can arise from support size and spatial sampling density (compare Appendix A). Figure B4c shows a large "loss" of

downward trends/"gain" in upward trends for CAR when switching from grids to basins. This difference propagates into the hotspot metric (Figure B4d), where CAR shows a >10 percentage-point lower share of persistent or emerging hotspots under HYBAS aggregation. Overall, the close grid–HYBAS agreement in Figure B4d indicates that grid-based hotspot diagnostics are robust at global and broad regional scales, while larger deviations primarily occur in regions where hotspots are more transitional (i.e., where a larger share of hotspot units is classified as stabilizing). In these regions, grid-based aggregation can yield slightly higher persistent/emerging shares than HYBAS.

B.5 Supporting trend results

Table B1: IPCC-region-specific trend estimates combining 19- and 9-year windows used in the publication. For each region, the number of samples (N Samples), the share of units with significant trends ($|\text{slope}| \geq 0.1$ m yr^{-1}; % Trend (Abs)), and the relative frequencies of significant upward (% Trend Up) and downward (% Trend Down) trends are listed.

Region	N Samples	% Trend (Abs)	% Trend Up	% Trend Down
ARO	1	0	0	0
CAR	165	15	52	48
CAU	146	32	62	38
CNA	504	64	74	26
EAS	16	56	44	56
EAU	1,922	41	59	41
ENA	3,532	19	43	57
ESAF	43	53	35	65
MDG	22	18	75	25
MED	1,924	34	41	59
NAU	1,303	40	82	18
NCA	1,377	44	44	56
NEN	3	0	0	0
NES	42	50	62	38
NEU	9,403	9	69	31
NPO	15	0	0	0
NSA	12	0	0	0
NWN	5	0	0	0
NZ	830	17	44	56
SAS	1,516	27	58	42
SAU	4,325	29	18	82
SCA	418	35	40	60
SES	23	9	0	100
WCE	4,509	8	63	37
WNA	4,079	45	35	65
WSAF	339	49	16	84

Table B2: IPCC-region-specific 9-year trend estimates. For each region, the number of samples (N Samples), the share of units with significant trends ($|\text{slope}| \geq 0.1$ m yr^{-1}; % Trend (Abs)), and the relative frequencies of significant upward (% Trend Up) and downward (% Trend Down) trends are listed.

Region	**N Samples**	**% Trend (Abs)**	**% Trend Up**	**% Trend Down**
ARO	1	0	0	0
CAR	165	16	48	52
CAU	146	36	63	37
CNA	504	62	76	24
EAS	16	56	44	56
EAU	1,922	53	23	77
ENA	3,532	26	40	60
ESAF	43	63	33	67
MDG	22	50	55	45
MED	1,924	38	39	61
NAU	1,303	48	82	18
NCA	1,377	45	45	55
NEN	3	0	0	0
NES	42	50	62	38
NEU	9,403	11	64	36
NPO	15	7	100	0
NSA	12	0	0	0
NWN	5	0	0	0
NZ	830	21	42	58
SAS	1,516	34	30	70
SAU	4,325	33	24	76
SCA	418	36	36	64
SES	23	9	0	100
WCE	4,509	10	52	48
WNA	4,079	46	39	61
WSAF	339	53	14	86

Table B3: IPCC-region-specific 19-year trend estimates. For each region, the number of samples (N Samples), the share of units with significant trends (|slope| $\geq$ 0.1 m yr^{-1}; % Trend (Abs)), and the relative frequencies of significant upward (% Trend Up) and downward (% Trend Down) trends are listed.

Region	**N Samples**	**% Trend (Abs)**	**% Trend Up**	**% Trend Down**
ARO	–	–	–	–
CAR	96	14	54	46
CAU	43	33	43	57
CNA	247	66	75	25
EAS	–	–	–	–
EAU	1,178	35	91	9
ENA	2,254	18	48	52
ESAF	17	53	22	78
MDG	13	15	50	50
MED	995	30	38	62
NAU	854	39	86	14
NCA	257	44	41	59
NEN	–	–	–	–
NES	–	–	–	–
NEU	4,022	5	70	30
NPO	10	0	0	0
NSA	–	–	–	–
NWN	2	0	0	0
NZ	484	12	43	57
SAS	1,044	22	77	23
SAU	2,821	23	15	85
SCA	42	40	65	35
SES	–	–	–	–
WCE	1,700	7	66	34
WNA	1,501	47	23	77
WSAF	81	38	23	77

Appendix C Disentangling coastal groundwater level dynamics in a global dataset

C.1 Time series outlier detection with DBSCAN

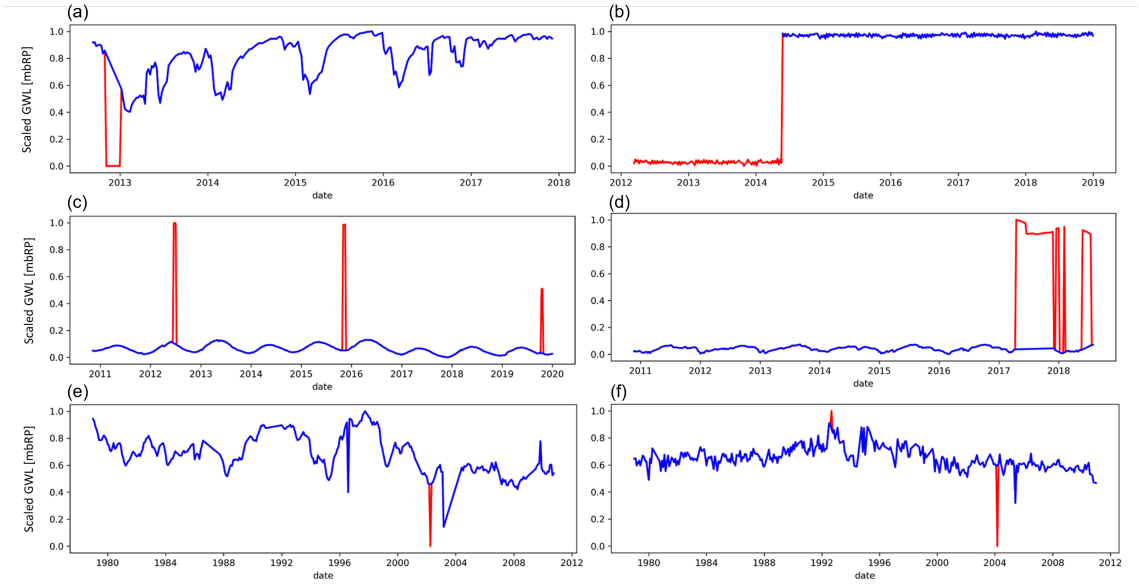


Figure C1: Examples of GWL time series outlier removal with DBSCAN (scikit-learn library in Python): raw (scaled) time series are shown in blue, and the anomalies identified by DBSCAN (Ester et al. 1996) are shown in red. The use of DBSCAN is less successful or difficult to verify in the bottom two examples (in panel e, the first abrupt change in direction of the GWL between 1996 and 2000 is not identified as an outlier, although one would expect one based on the magnitude and position of the change; in panel f, from visual inspection, it is unclear whether the identified outliers in both directions are due to measurement error or human activity or reflect natural behavior of the groundwater system, e.g., extreme events). In summary, like other tested outlier detection methods, DBSCAN does not allow us to detect all types of outliers and anomalies that we would expect should be removed to represent undisturbed GWL dynamics. With the parameters set, we can successfully detect most density-dependent outliers, but many values were incorrectly flagged as outliers (false positives). Therefore, we performed a visual inspection of all time series where DBSCAN identified potential outliers.

C.2 Indices

Table C1: List of GWL dynamic typology and indices used in this study. Differing names in brackets refer to deviating abbreviations in Heudorfer et al. (2019).

Aspect	**Component**	**Index names**
Distribution	boundness	fdc.slope_0.8_1 (dc.slp.u); fdc.slope_0_0.1 (dc.slp.l); l1; l3; median (med); peakts.avg
Distribution	density	colwells.C (colwell.C); fdc.range_0.1_0.9 (dc.rng.01.09); fdc.range_0.2_0.8 (dc.rng.02.08); fdc.range_0.25_0.75 (dc.rng.025.075); fdc.slope_0.3_0.7 (dc.slp.m); l2; l4; meanann.max (avg.ann.max)
Distribution	modality	bimodality (bimod); dip; silverman (bandwd)
Shape	scale	peakbasetime.avg (peakbase.avg); pulse.dur.high (pulse.dur.h); pulse.dur.low (pulse.dur.l); sqr.avg
Shape	slope	fall.avg; fall.cv; reces.const; recov.const; rise.avg; rise.cv
Structure	amplitude	amp.max
Structure	flashiness	baker2 (baker); BFI; Lyapunov; pathlength2; reversals.avg (rev.avg); reversals.cv (rev.cv)
Structure	interannual variation	autocorrelation (autocor); BFS; Hurst; intannfluc.y (iaf.y); pulse.count.high (pulse.count.h); pulse.count.low (pulse.count.l)
Structure	seasonality-magnitude	cvmon.min; intannfluc.s (iaf.s); parde.seasonality (parde)
Structure	seasonality-timing	colwells.M; varjulday.max (newly introduced instead of varjuld.min: CV of date of annual maximum head)

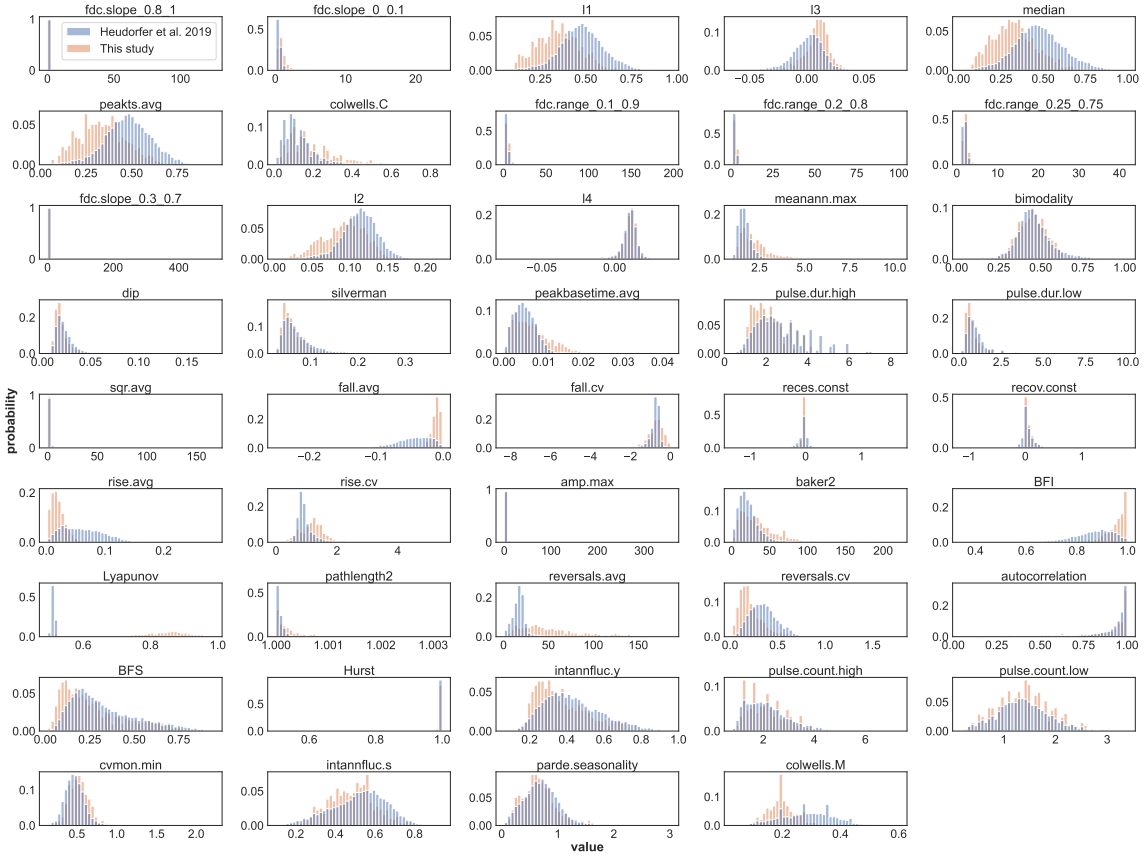


Figure C2: Comparison of the value distributions of the groundwater hydrograph indices calculated in this study (Table C1) with the values published by Heudorfer et al. (2019) (dataset: Haaf and Heudorfer 2018). The magnitude and variability of values for many indices are similar. Shifted or significantly different ranges of values are observed for some indices. Such differences exist for indices of different aspects of the hydrograph (structure, distribution, and shape).

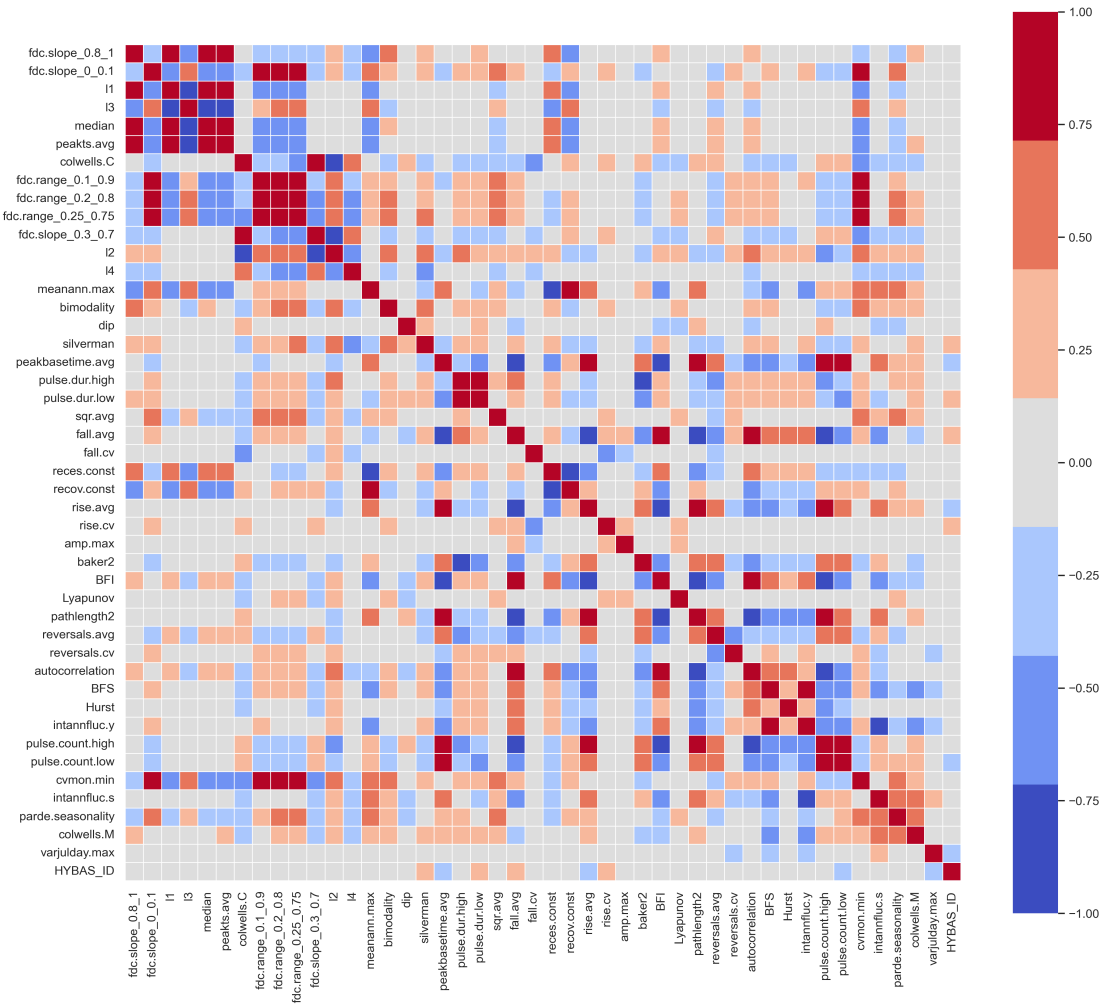


Figure C3: Correlation matrix of the index values calculated in this study.

C.3 Clustering

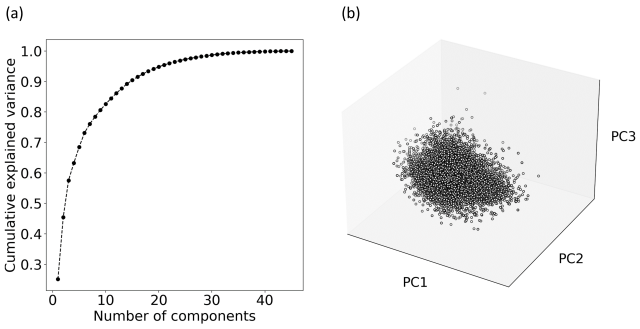


Figure C4: PCA with the index dataset. (a) Cumulative explained variance with an increasing number of principal components (PCs). (b) Dataset value space of the first three PCs (which together explain more than 50% of the variance of the dataset).

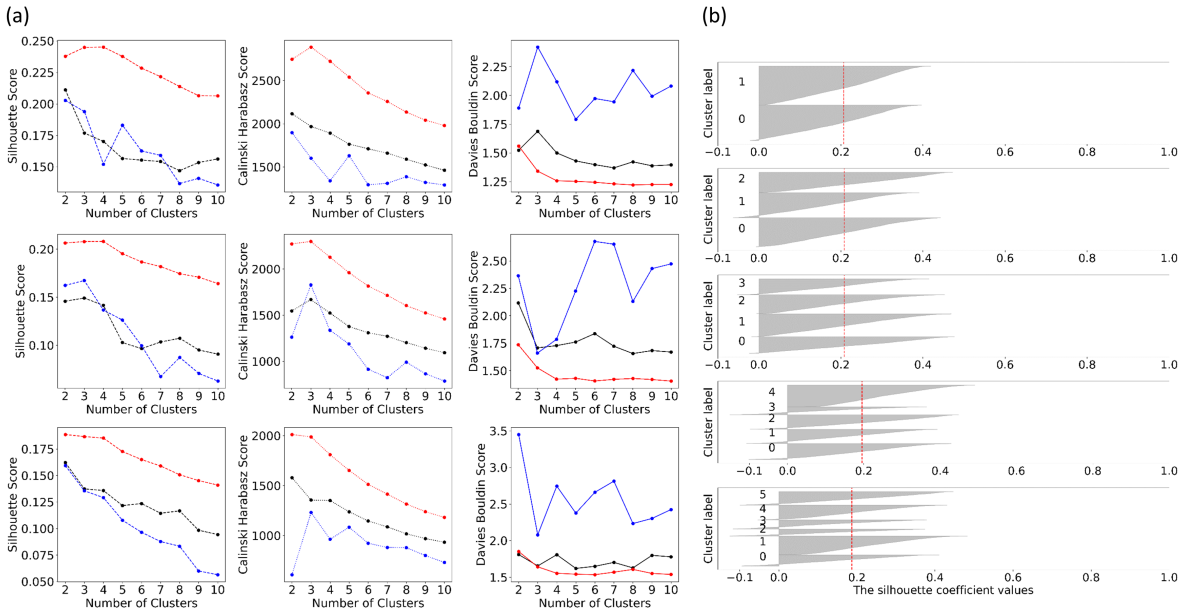


Figure C5: Evaluation metrics – Silhouette (Rousseeuw 1987), Calinski-Harabasz (Caliński and Harabasz 1974) and Davies-Bouldin (Davies and Bouldin 1979) – used to find the best cluster separation. The Silhouette score ranges from -1 to 1, where a higher value indicates better-defined clusters. Better clustering results are also indicated by higher values for the Calinski-Harabasz score and by lower values for the Davies-Bouldin score. (a) Metrics are shown for different proportions of explained variance represented by PCs (top to bottom row: 60%, 70%, and 80%) and for different clustering algorithms (k-means clustering, red; Gaussian mixture, blue; and hierarchical clustering, black). (b) Silhouette scores for various clusters using the k-means algorithm and PCs representing 70% of explained variance.

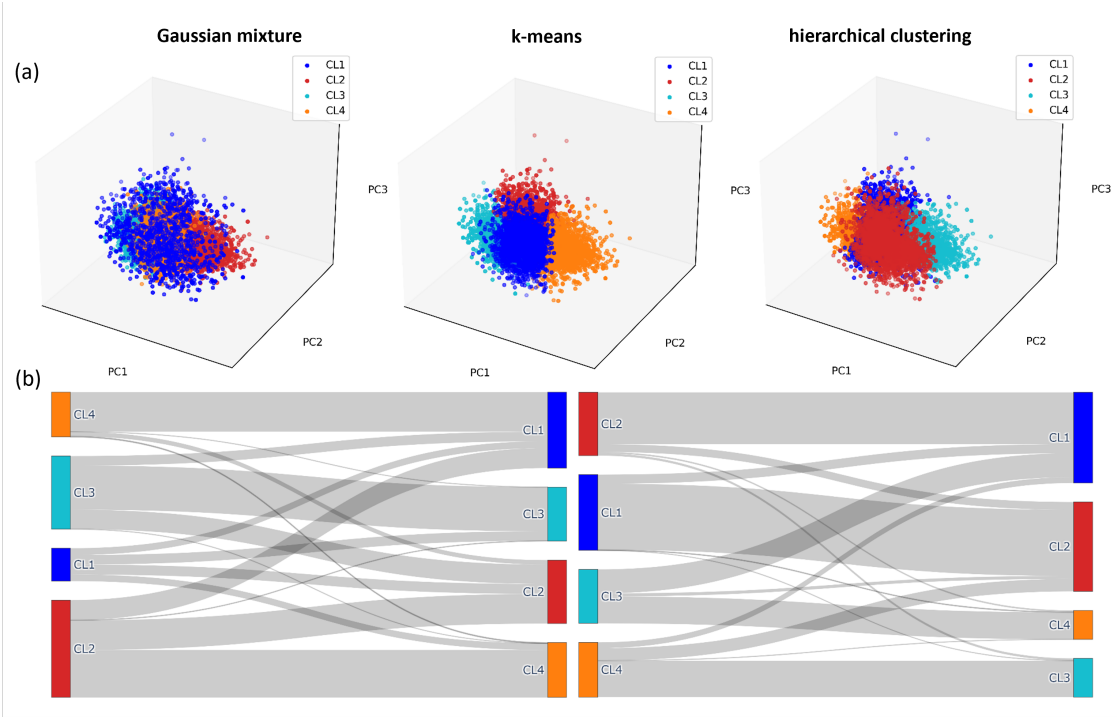


Figure C6: Assignment of the time series to four different clusters by three algorithms (Gaussian mixture, k-means clustering, and hierarchical clustering). (a) Illustrative cluster assignment within the range of values of the first three PCs of the dataset. (b) Sankey plot allows comparison of the quantitative distribution of the samples among the clusters (left between Gaussian mixture and k-means and right between k-means and hierarchical clustering).

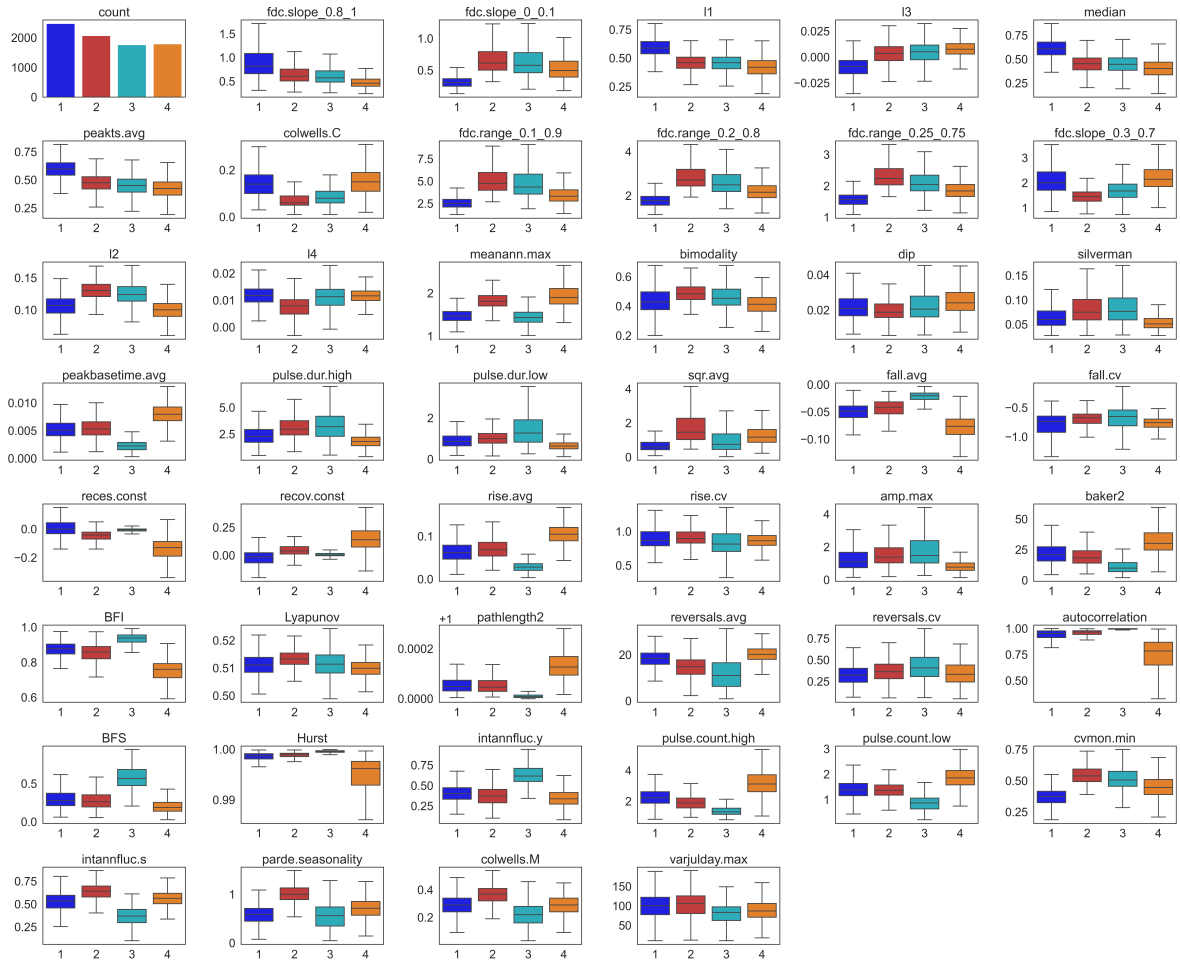


Figure C7: Within-cluster variability of all indices used for clustering.

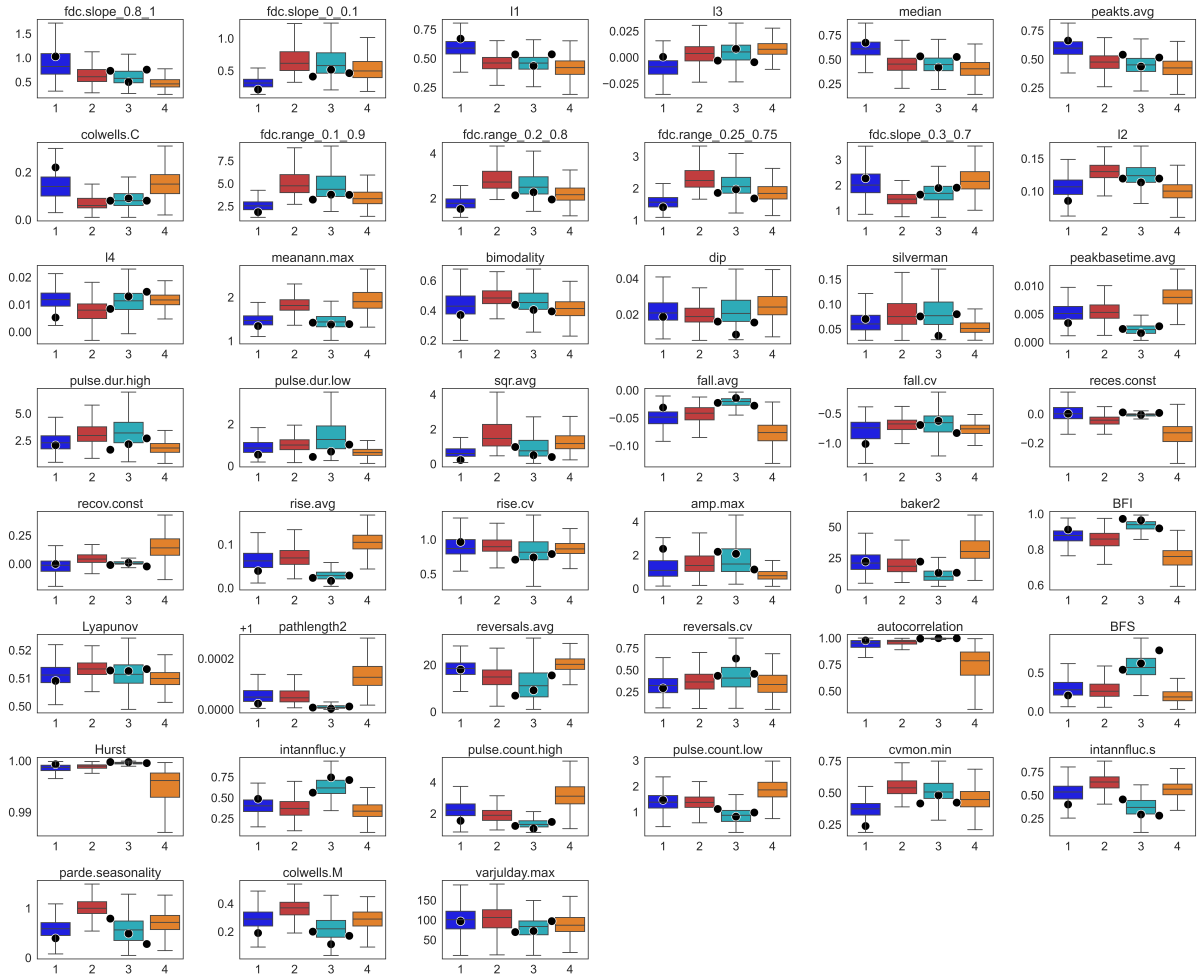


Figure C8: Within-cluster variability of all indices used for clustering (Figure C7) plotted with index values (dots) of the four example wells from the case study region in northern Germany (Figure 12a). The points drawn between CL2 and CL3 and between CL4 and CL3 are wells with the highest probability for CL2 and CL4 in the random forest model but belong to CL3 according to the clustering.

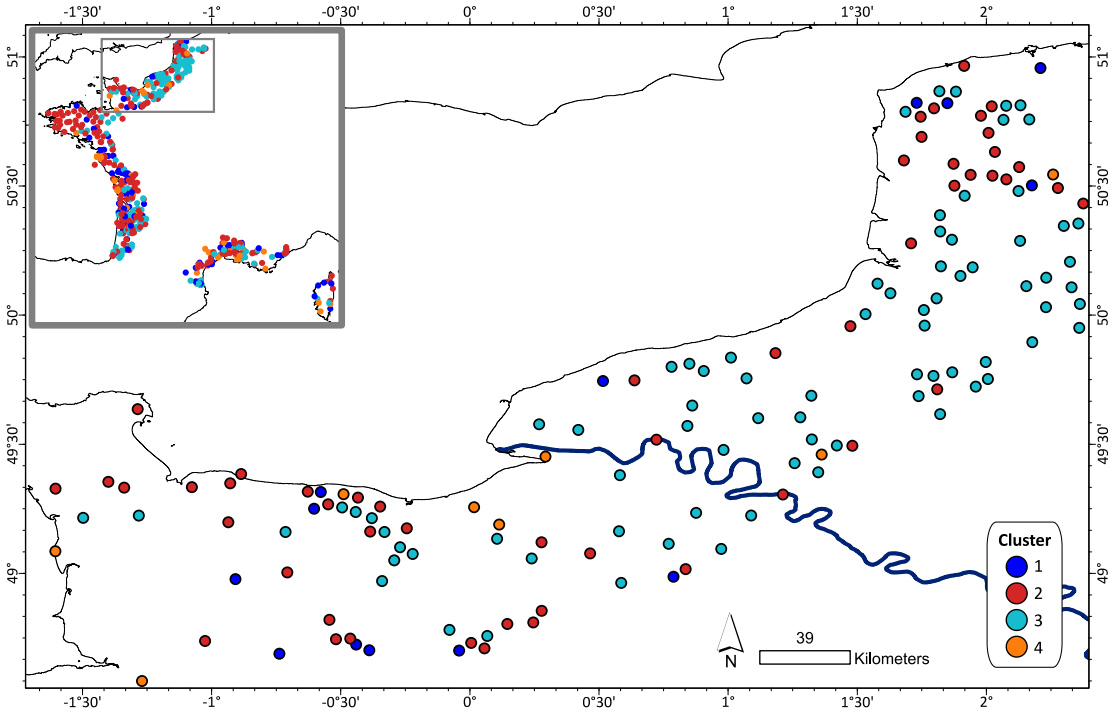


Figure C9: Assignment of clusters to well locations in the Normandy, Paris Basin, displayed with the Seine River. Overlapping well markers were jittered at a minimum spacing of 1000 meters and thus no longer represent the original well locations. The window in the upper left corner shows the Normandy region in the north of France.

C.4 Random forest classification

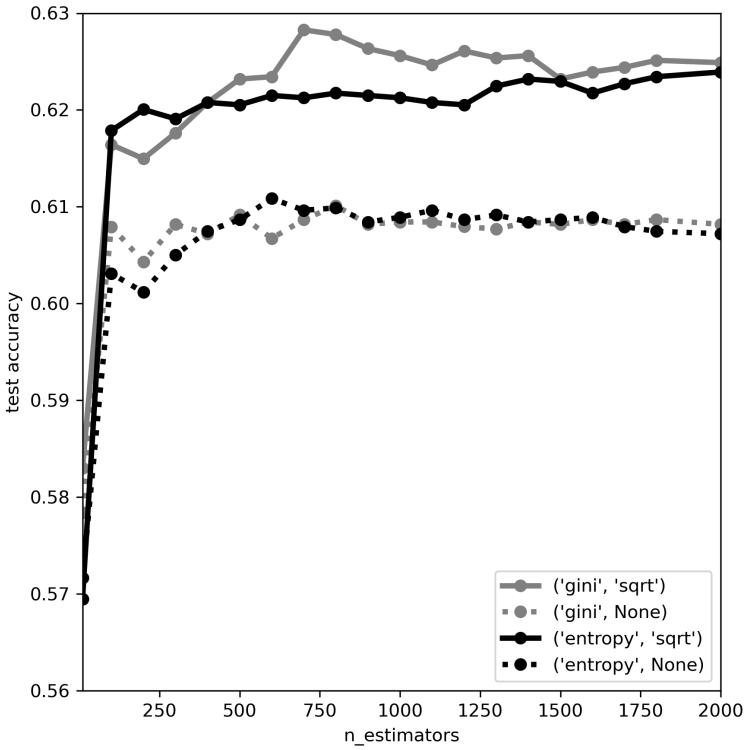


Figure C10: Random forest hyperparameter tuning within a 5-fold cross-validation framework using the training data.

C.5 SHAP

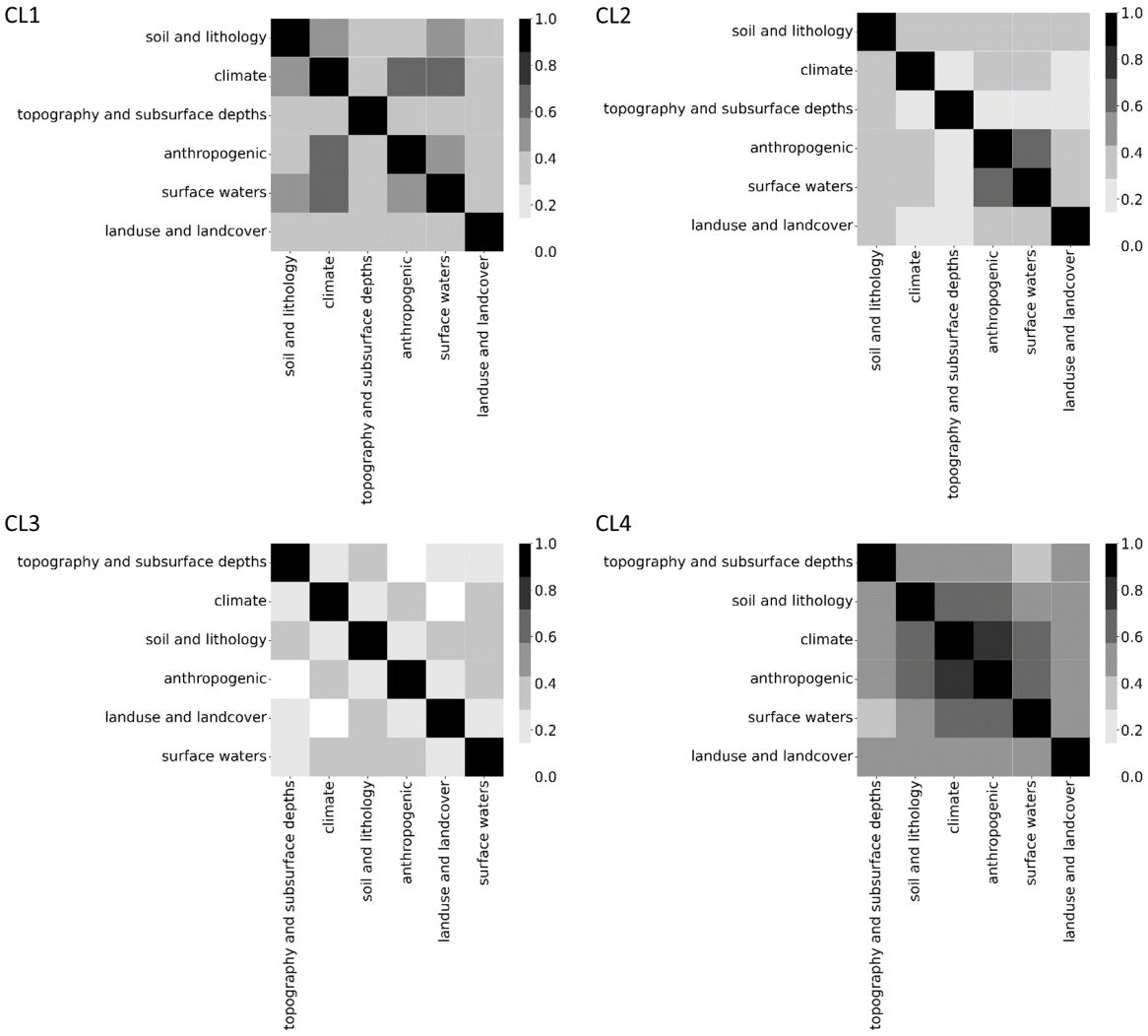


Figure C11: Absolute Spearman correlation of SHAP values for aggregated features (SHAP values added up for features from the same attribute categories from Table 4).

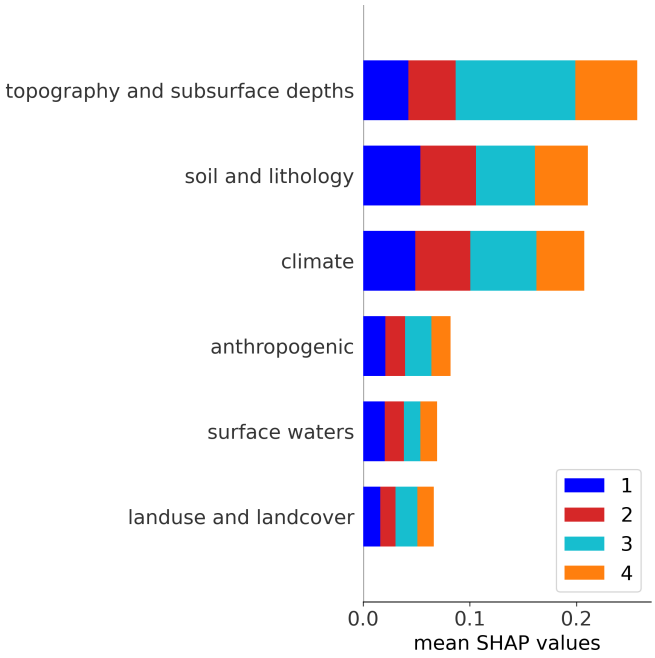


Figure C12: Overall SHAP feature importance stacked for individual clusters for aggregated features (SHAP values added up for features from the same attribute categories from Table 4).

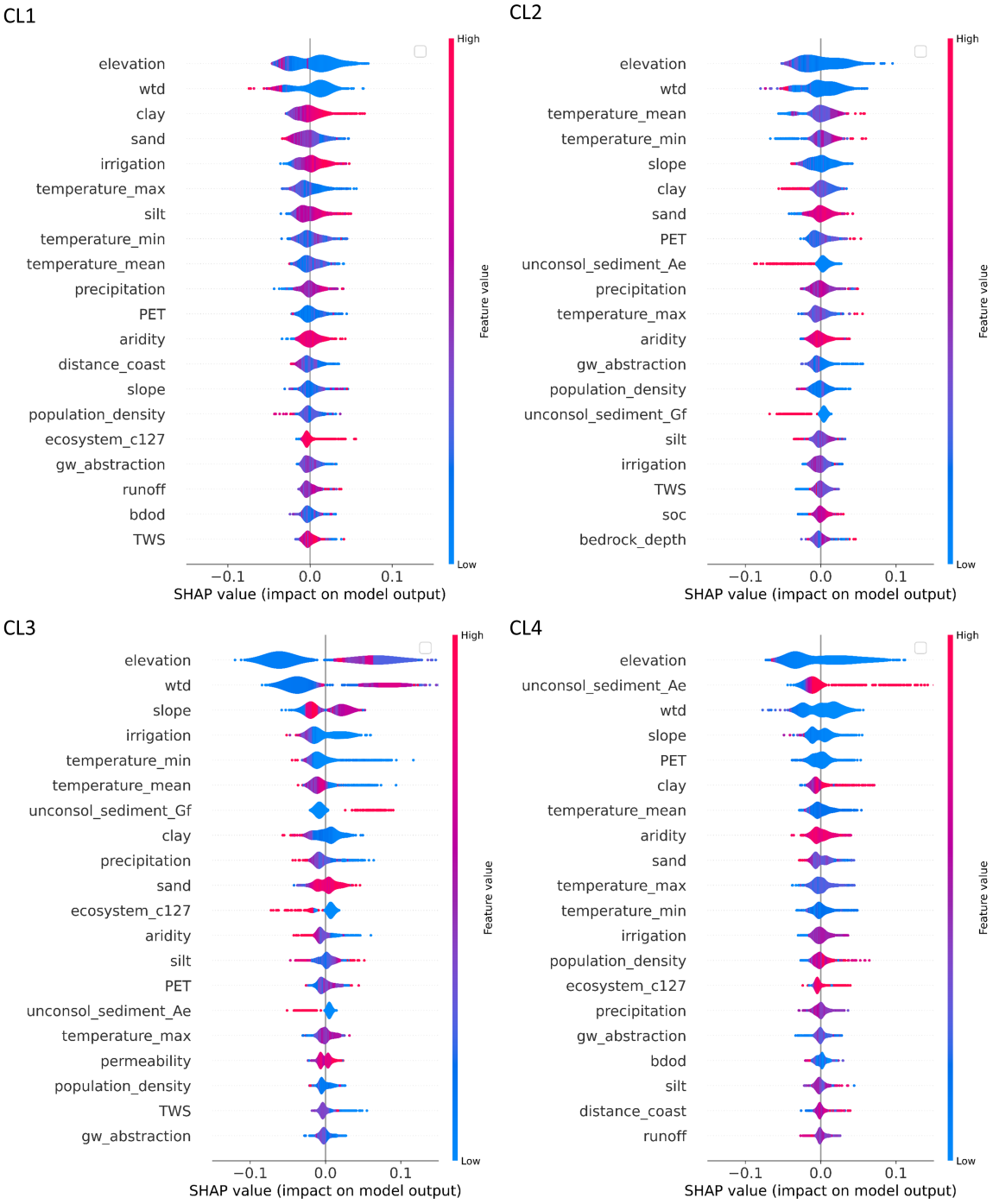


Figure C13: Violin plots show the distribution of SHAP feature importance and feature effect for the 20 highest-ranked features of each cluster. SHAP values describe the impact of a given feature on the prediction of the model for a given cluster (the prediction of "true" for the selected cluster). A red color corresponds to high feature values, and a blue color corresponds to low feature values (for one-hot-encoded features, a value of 0 (blue) corresponds to the presence of the feature class, and a value of 1 (red) corresponds to the absence of the feature class). Overlapping instances widen the violin shape in the direction of the y-axis, and the violin at that position is colored according to the average feature value.

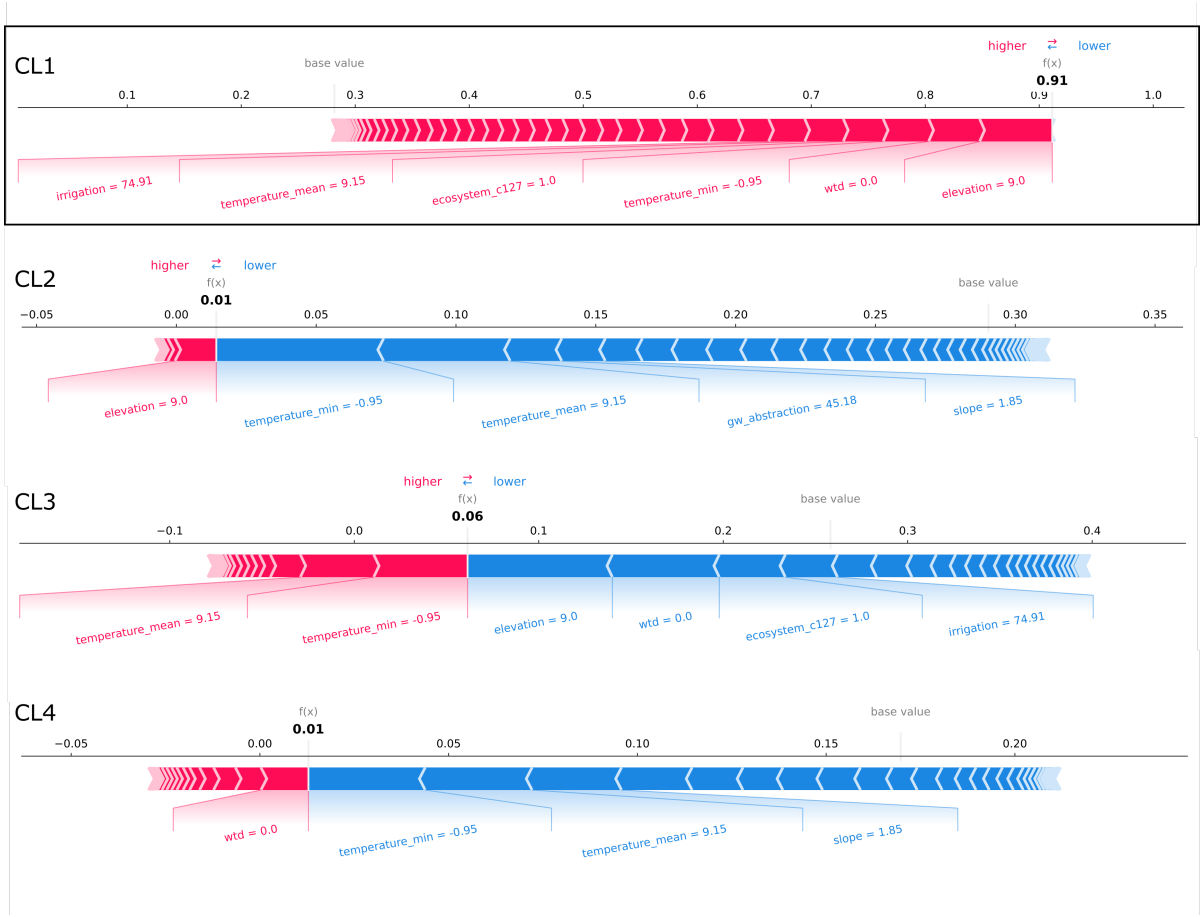


Figure C14: SHAP force plots explaining the correct prediction of the CL1 exemplary well site from the case study region in northern Germany (Figure 12a). At this well site, the probability that CL1 is predicted is the largest in the case study (test dataset). The average scores of all classifications made by the random forest model for the training dataset are 0.28, 0.29, 0.26, and 0.17 for CL1–CL4, respectively (base values). The model's probability scores for predicting the clusters at the exemplary well site (bold printed numbers) sum up to 100%. Features that are important for the respective predictions are displayed with their values. Their importance and effect can be assessed by the SHAP values visualized by bar length and color (i.e., the larger the feature's share of the bar the more important; red represents rejecting effect, and blue represents supporting effect).

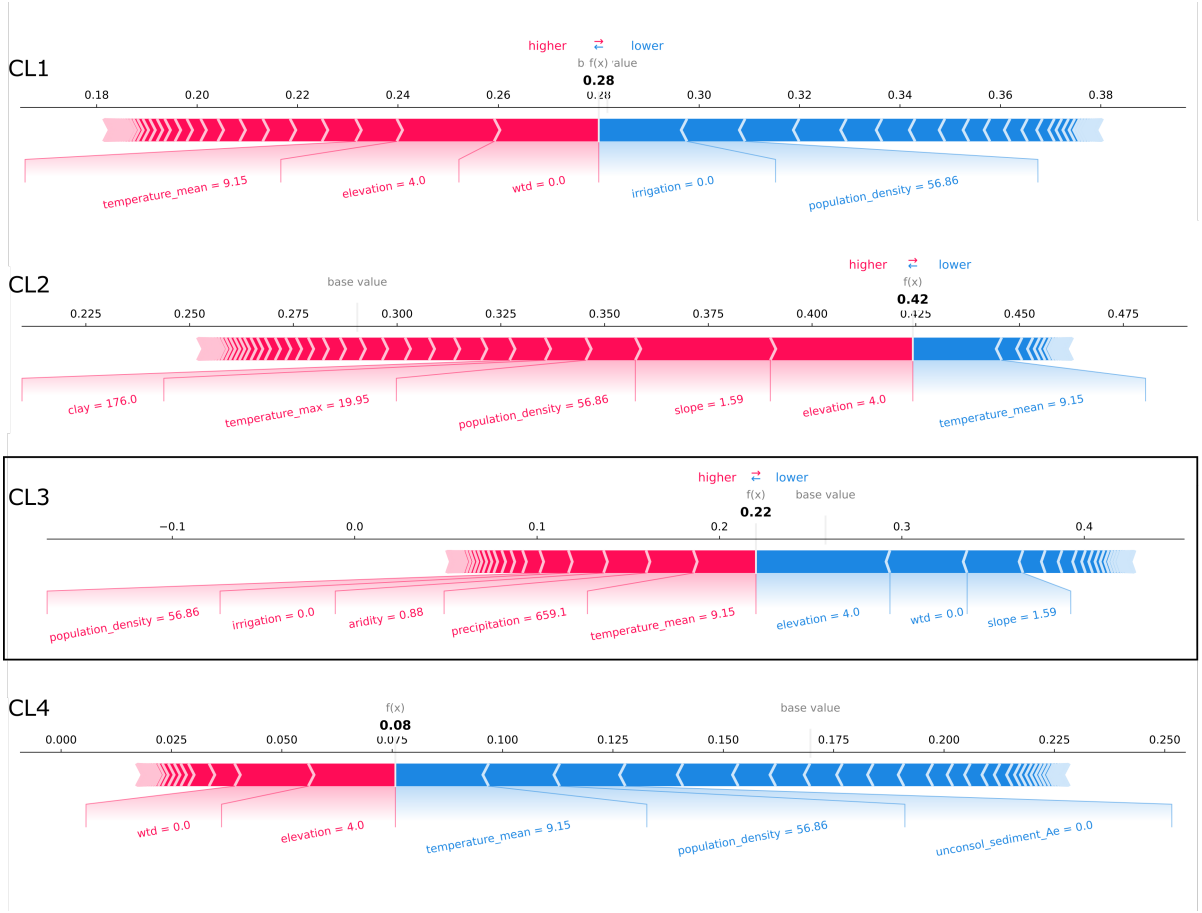


Figure C15: SHAP force plots explaining the prediction of the CL3 exemplary well site from the case study region in northern Germany (Figure 12a) that is confused with CL2. At this well site, the probability that CL2 is predicted is the largest in the case study (test dataset). The average scores of all classifications made by the random forest model for the training dataset are 0.28, 0.29, 0.26, and 0.17 for CL1–CL4, respectively (base values). The model's probability scores for predicting the clusters at the exemplary well site (bold printed numbers) sum up to 100%. Features that are important for the respective predictions are displayed with their values. Their importance and effect can be assessed by the SHAP values visualized by bar length and color (i.e., the larger the feature's share of the bar the more important; red represents rejecting effect, and blue represents supporting effect).



Figure C16: SHAP force plots explaining the correct prediction of the CL3 exemplary well site from the case study region in northern Germany (Figure 12a). At this well site, the probability that CL3 is predicted is the largest in the case study (test dataset). The average scores of all classifications made by the random forest model for the training dataset are 0.28, 0.29, 0.26, and 0.17 for CL1–CL4, respectively (base values). The model's probability scores for predicting the clusters at the exemplary well site (bold printed numbers) sum up to 100%. Features that are important for the respective predictions are displayed with their values. Their importance and effect can be assessed by the SHAP values visualized by bar length and color (i.e., the larger the feature's share of the bar the more important; red represents rejecting effect, and blue represents supporting effect).

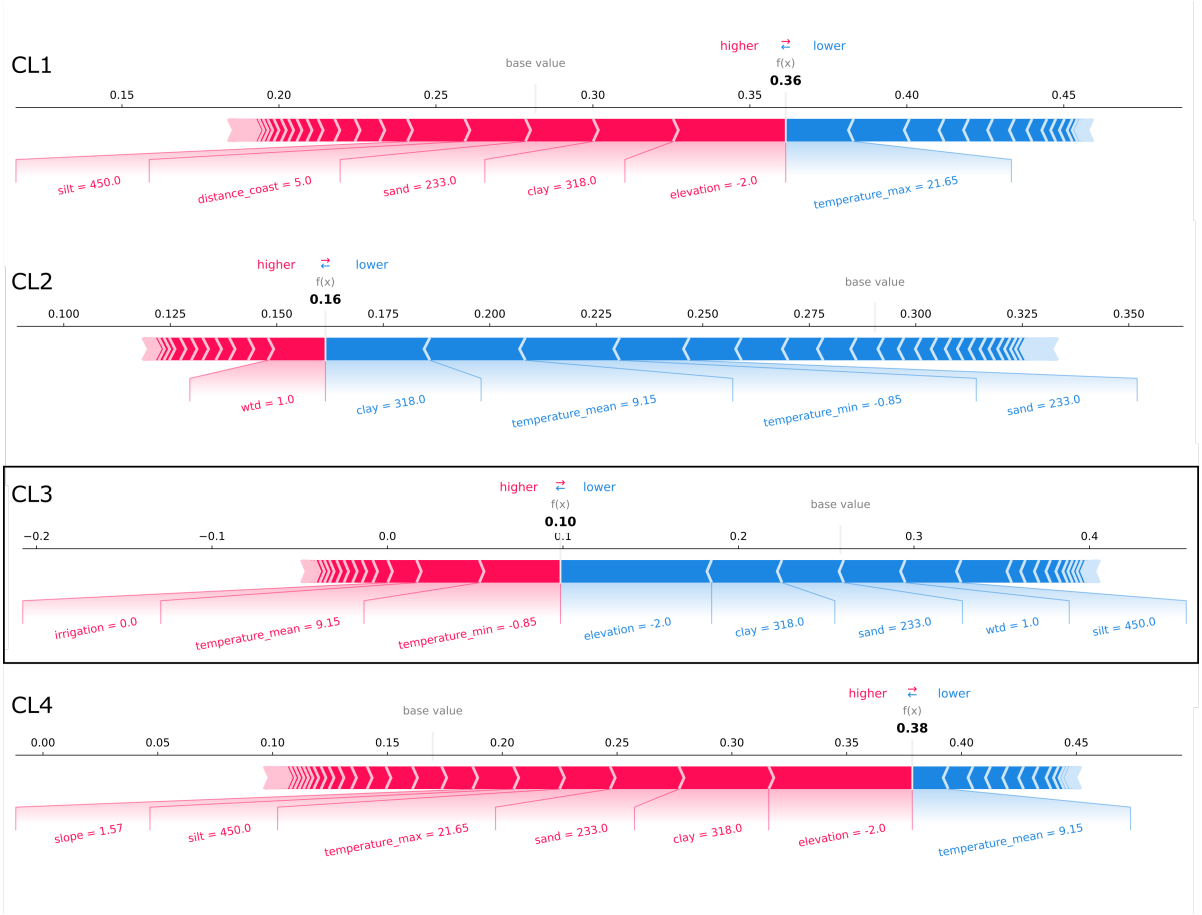


Figure C17: SHAP force plots explaining the prediction of the CL3 exemplary well site from the case study region in northern Germany (Figure 12a) that is confused with CL4. At this well site, the probability that CL4 is predicted is the largest in the case study (test dataset). The average scores of all classifications made by the random forest model for the training dataset are 0.28, 0.29, 0.26, and 0.17 for CL1–CL4, respectively (base values). The model's probability scores for predicting the clusters at the exemplary well site (bold printed numbers) sum up to 100%. Features that are important for the respective predictions are displayed with their values. Their importance and effect can be assessed by the SHAP values visualized by bar length and color (i.e., the larger the feature's share of the bar the more important; red represents rejecting effect, and blue represents supporting effect).

Appendix D Multi-site deep learning for groundwater level prediction across global datasets: toward scalable applications under data scarcity

D.1 Descriptive statistics of groundwater and climate data

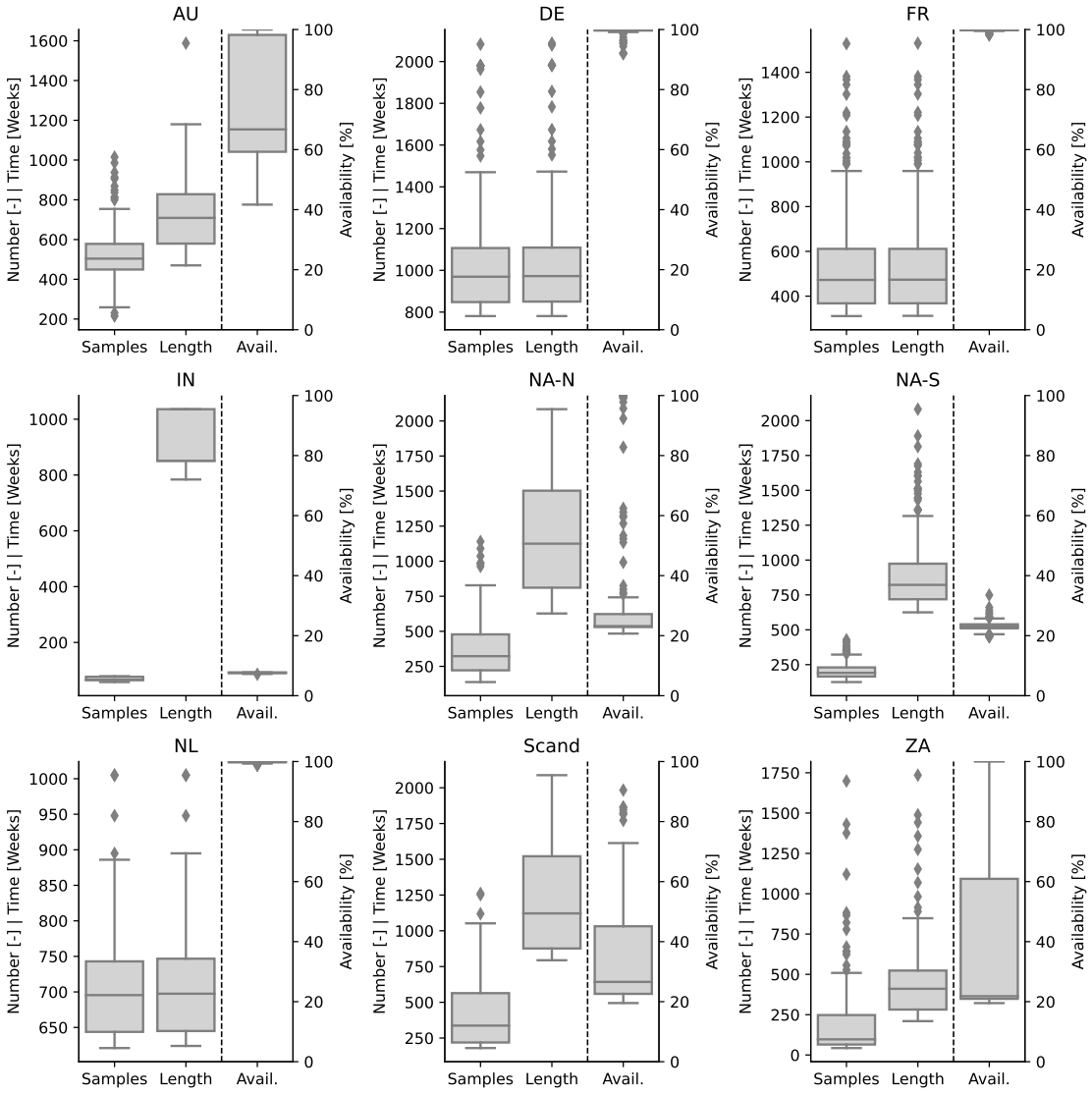


Figure D1: Boxplots showing GWL time series data quantity metrics for the nine focus regions: the distribution of absolute sample numbers (Samples), time series lengths (Length), and availability (Avail.).

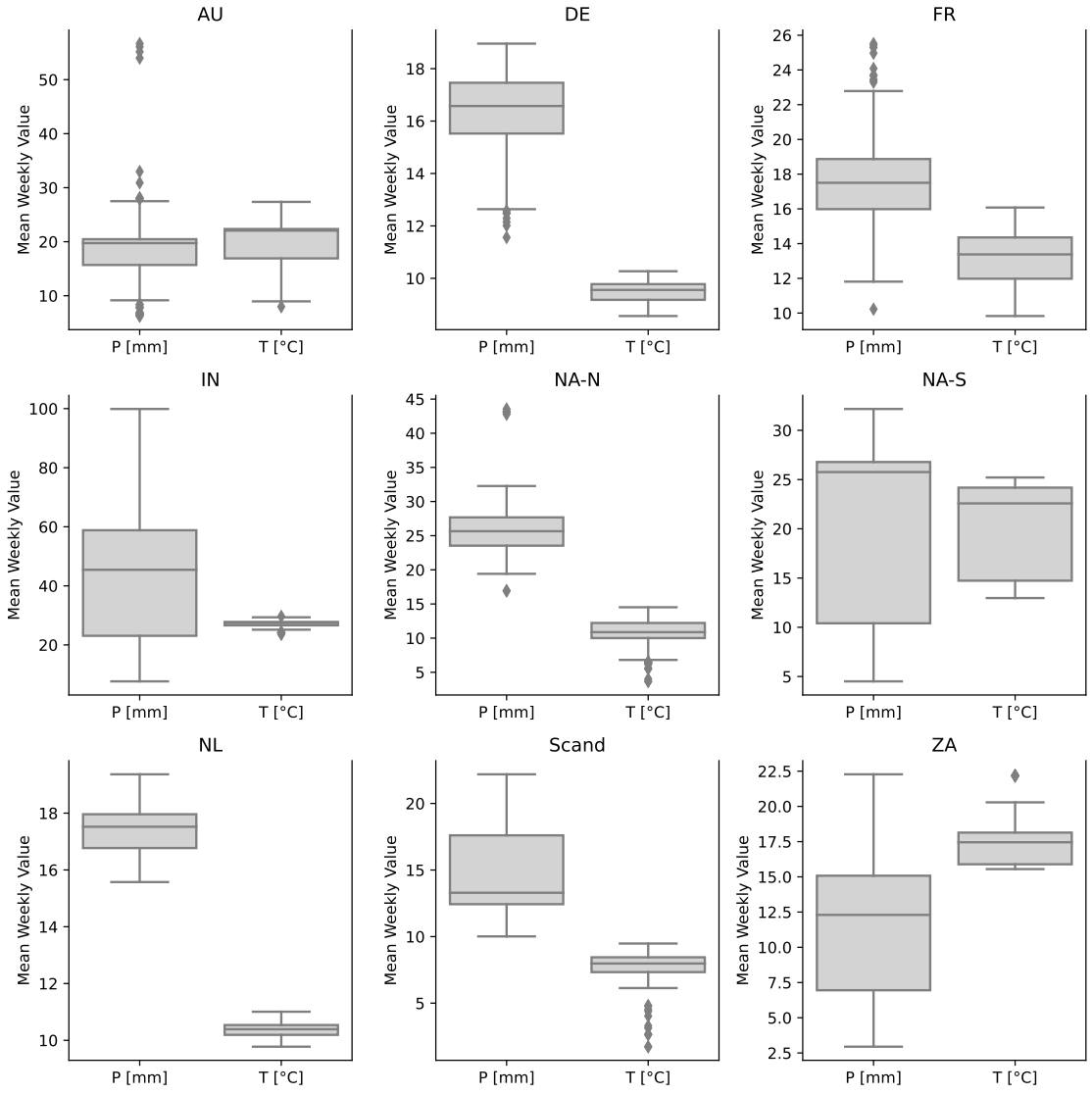


Figure D2: Boxplots showing mean weekly precipitation (P) and surface temperature (T) values for the well locations in the nine focus regions.

D.2 Model hyperparameters

Our LSTM model is built with the following hyperparameters from Kratzert et al. (2021) and Kratzert et al. (2024): hidden size (265); dropout (40%); optimizer (Adam); number of epochs (30); learning rate (1×10^{-3}, reduced to 5×10^{-4} at epoch 20, further reduced to 1×10^{-4} at epoch 25).

We chose a sequence length of 52 with weekly time steps which corresponds to one calendar year. This is a typical length chosen for rainfall-runoff modeling (Kratzert et al. 2021; Kratzert et al. 2024) and GWL modeling (Wunsch et al. 2022; Heudorfer et al. 2024).

For the loss function, we used the MSE masked for data gaps with standardized GWL data.

D.3 Model evaluation

Model performance during the test period was assessed using six standard evaluation metrics:

1. Nash-Sutcliffe Efficiency (NSE): Indicates how well the model predictions match observed data, with values closer to 1 indicating better performance.

2. Kling-Gupta Efficiency (KGE): Integrates correlation, variability, and bias to provide a comprehensive measure of model skill.

3. Coefficient of Determination (R^2): Reflects the strength of the linear relationship between predicted and observed values.

4. Bias: Measures the average difference between predicted and observed values, indicating systematic over- or underestimation.

5. Mean Squared Error (MSE): Represents the average squared prediction error, with lower values indicating higher accuracy.

6. Root Mean Squared Error (RMSE): The square root of MSE, providing an interpretable measure of average error magnitude in the same units as the data.

Table D1: Median model evaluation metrics across 10 random initializations for each model variant (*nofeat*, *embeddings*, *ohefeat*) and dataset modeled. Performance is assessed during the test period using six standard metrics: Nash–Sutcliffe Efficiency (NSE), Kling–Gupta Efficiency (KGE), coefficient of determination (R^2), bias, mean squared error (MSE), and root mean squared error (RMSE). Higher values of NSE, KGE, and R^2, and lower values of bias, MSE, and RMSE indicate better model performance.

Region	**Model**	**NSE**	**KGE**	R^2	**Bias**	**MSE**	**RMSE**
AU	*nofeat*	0.499	0.462	0.567	-0.013	0.290	0.538
	embeddings	0.696	0.617	0.679	-0.055	0.186	0.431
	ohefeat	0.670	0.602	0.646	-0.060	0.190	0.436
DE	*nofeat*	0.475	0.602	0.584	0.007	0.044	0.210
	embeddings	0.729	0.794	0.824	-0.003	0.021	0.145
	ohefeat	0.699	0.771	0.813	0.000	0.025	0.158
FR	*nofeat*	0.602	0.617	0.675	0.031	0.243	0.493
	embeddings	0.787	0.803	0.864	-0.019	0.129	0.360
	ohefeat	0.794	0.796	0.862	-0.010	0.130	0.360
IN	*nofeat*	0.510	0.615	0.636	-0.195	0.697	0.835
	embeddings	0.523	0.634	0.650	-0.176	0.667	0.817
	ohefeat	0.470	0.605	0.617	-0.195	0.743	0.862
NA-N	*nofeat*	0.536	0.606	0.606	-0.026	0.090	0.300
	embeddings	0.694	0.776	0.745	-0.025	0.057	0.239
	ohefeat	0.682	0.774	0.729	-0.021	0.064	0.254
NA-S	*nofeat*	0.398	0.498	0.530	-0.038	0.200	0.447
	embeddings	0.461	0.588	0.558	-0.019	0.189	0.435
	ohefeat	0.406	0.551	0.521	-0.008	0.214	0.462
NL	*nofeat*	0.483	0.543	0.568	0.015	0.020	0.141
	embeddings	0.685	0.764	0.769	0.005	0.012	0.112
	ohefeat	0.677	0.740	0.752	0.008	0.013	0.112
Scand	*nofeat*	0.390	0.505	0.442	-0.031	0.128	0.357
	embeddings	0.586	0.660	0.622	-0.027	0.084	0.290
	ohefeat	0.493	0.629	0.595	-0.029	0.101	0.318
ZA	*nofeat*	0.395	0.441	0.536	-0.010	0.214	0.462
	embeddings	0.593	0.581	0.657	-0.014	0.124	0.352
	ohefeat	0.554	0.562	0.611	-0.012	0.135	0.368
Global	*nofeat*	0.443	0.525	0.546	-0.008	0.138	0.371
	embeddings	0.631	0.705	0.709	-0.021	0.088	0.297

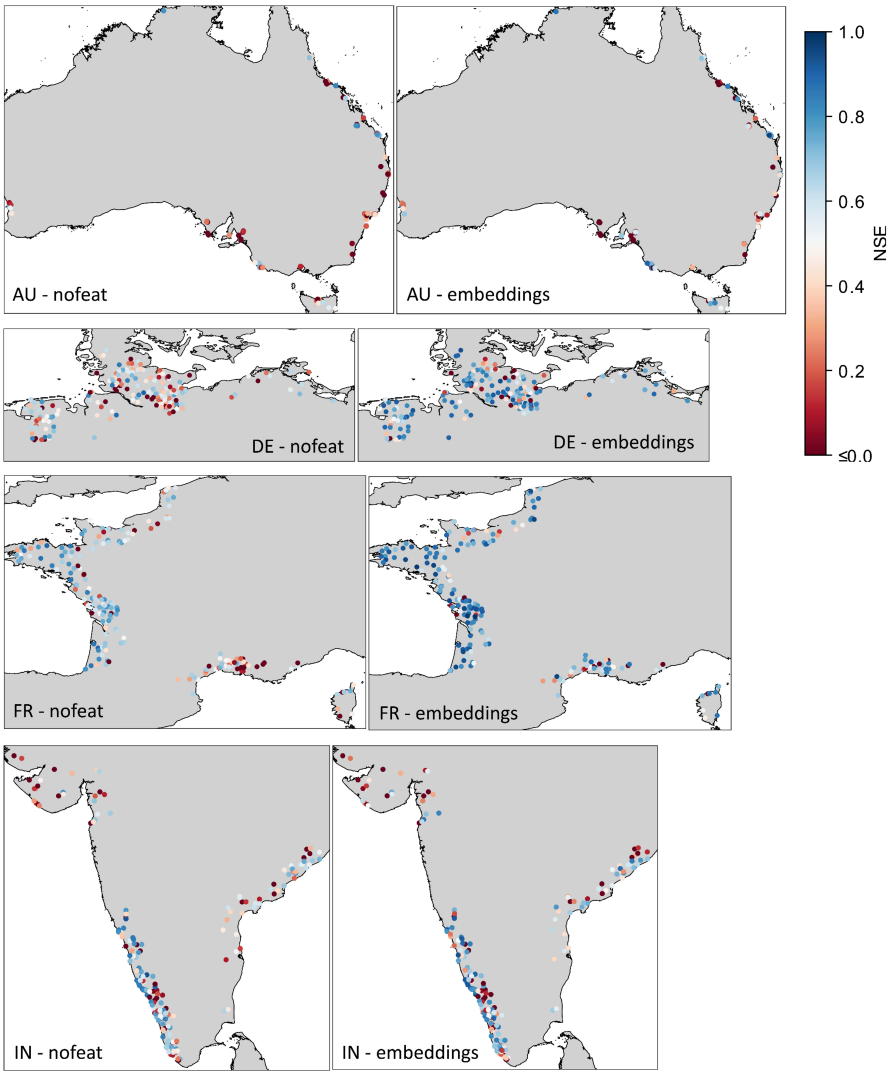


Figure D3: Spatial distribution of well-level predictive performance (median Nash–Sutcliffe efficiency, NSE) across selected focus regions. For each region, panels compare the *nofeat* model (left) and the *embeddings* model (right); values are the median across 10 random initializations. Colors indicate NSE (blue = higher, red = lower).

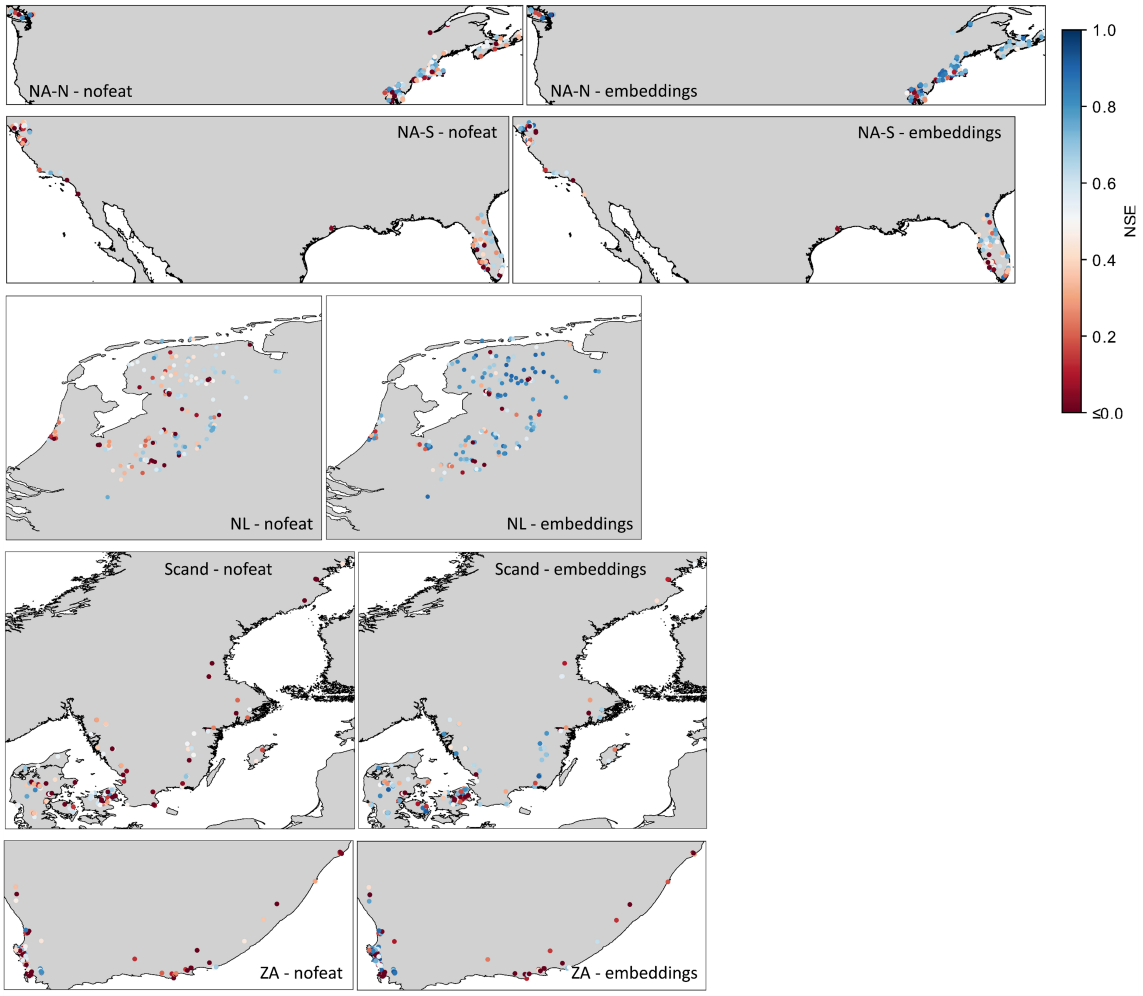


Figure D3 continued.

Figure D4 shows training and validation across seeds for each region and model variant. Loss decreases sharply in the first epochs and then approaches a plateau. The epoch of the minimum validation loss varies by seed and region. Training and validation curves remain close without sustained divergence, which would indicate overfitting. Model selection uses the global minimum validation loss within 30 epochs, and all reported test metrics (including NSE) are computed from that checkpoint on the unseen test period.

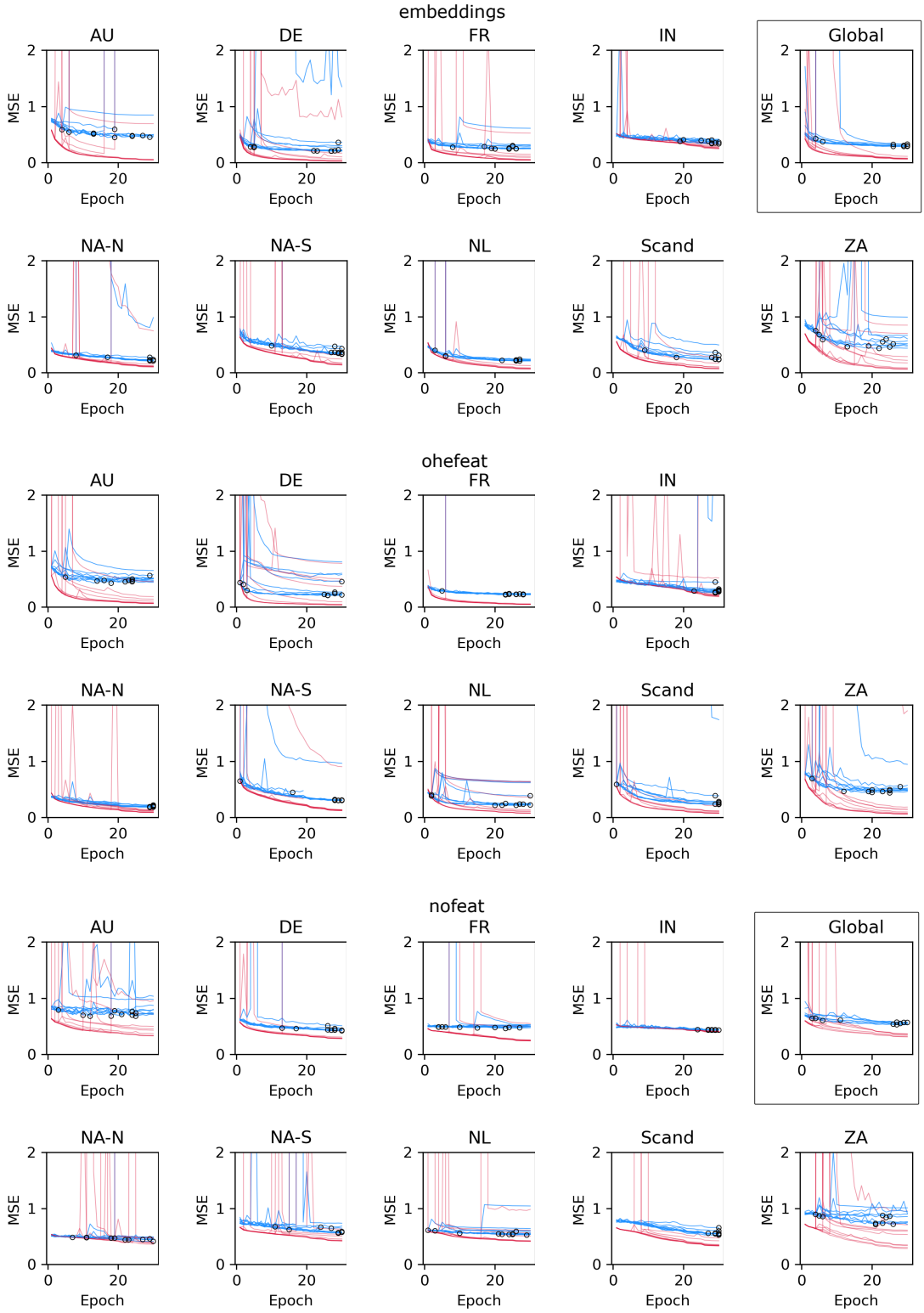


Figure D4: Learning curves of mean squared error (MSE), used as the training loss, for all regions (AU, DE, FR, IN, NA-N, NA-S, NL, Scand, ZA, Global) and model variants (*embeddings*, *ohefeat*, *nofeat*). Red lines show training loss; blue lines show validation loss; each seed contributes one pair of curves. Black dots mark the epoch of minimum validation loss, which defines the checkpoint used for test evaluation.

Use of Electronic Tools

The following electronic tools were used in preparing this dissertation: Overleaf/LaTeX (layout and typesetting), Microsoft Word (manuscript preparation), Citavi (reference management), Python, R, ArcGIS, and Microsoft Excel (data processing, analysis, modeling, and visualization), and ChatGPT (revision/polishing and initial suggestions for textual formulations and Python code).

Eidesstattliche Erklärung | Declaration on Oath

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

—

I hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated.

If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.

—

Ort, den | City, date — Unterschrift | Signature

Hamburg, den 19. Januar 2026