



STARS 4 Water

Upscaling machine learning models to other river basins

Deliverable 5.3

September 2026

Upscaling machine learning models to other river basins

Lead beneficiary	Juelich Research Centre – FZJ
Lead author(s)	Leandro Avila and Stefan Kollet
Work Package	5: Upscale and uptake for preparedness
Due date	25 September 2026
Submission date	28 September 2026
Contributors	Victor Gomez (UCM), Pedro Martinez (UCM), Helen Baron (UKCEH), Kolbjørn Engeland (NVE), Hege Hisdal (NVE), Trine Hegdahl (NVE), Maria-Helen Ramos (INRAE), Harm Duel (Deltares)

Dissemination Level		
PU	Public	X
SEN	Confidential, only for members of the consortium and the granting authority (including other EU institutions and bodies)	
CI	Classified, as referred to EU Decision 2015/444 and its implementing rules	

Version log			
Version	Date	Released by	Nature of Change
0.1	07.07.2026	Leandro Avila	Outline
0.7	24.08.2026	Leandro Avila	First draft
0.8	25.09.2026	Maria-Helena Ramos	Review
0.9	28.09.2026	Leandro Avila	Final version
1.0	28.09.2026	Harm Duel	Accepted

Citation

Avila, L., Kollet, S., Gomez, V., Martinez, P., Baron, H., Engeland, K., Hisdal, H., Hegdahl, T., Ramos, M.H. & Duel, H. (2026): Upscaling machine learning models to other river basins. Horizon Europe Project STARS4Water. Deliverable D5.3. Grant Agreement No 101059372



The STARS4Water project has received funding from the European Union's Horizon Europe research and innovation program under the Grant Agreement No 101059372

Disclaimer

The content of this deliverable does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s).

Summary

This report presents the results obtained from Task 5.2 (WP5), which focuses on upscaling machine learning models to other river basins and regions. The central questions addressed are whether data-driven models, initially trained at the river basin scale, retain sufficient predictive skill when transferred to different hydroclimatic regions and spatial domains, and whether they can serve as practical, computationally affordable tools for tracking key hydrological variables beyond the original river basin hubs.

The report evaluates four different machine learning approaches covering water availability and water quality: two ConvLSTM deep learning surrogate models, with one for daily Snow Water Equivalent (SWE) and another one for daily Water Table Depth (WTD), transferred from European training domains to the Rocky Mountains and the contiguous United States, respectively; a predictive mapping tool (MLMapper 2.0) based on ensemble tree algorithms, applied to identify groundwater-dependent ecosystems in central Chile and to assess groundwater exploitation suitability in southern Madagascar; and an ensemble-tree model for monthly reservoir storage forecasting, originally developed for European reservoirs and extended to tropical reservoirs in Thailand, evaluated through single-reservoir, regional, and cross-continental training configurations.

The clearest finding across all components is that models trained on high-resolution, physically based simulations or local observations can be transferred to different river basins and spatial domains without significant loss of predictive skill, though the degree of retraining or fine-tuning required depends on how similar the target domain is to the original training domain. This report presents results on the following data-driven models:

Across the four case studies, the results can be summarized as follows:

SWE: The ConvLSTM surrogate achieved median KGE values above 0.6 across the Rocky Mountains after fine-tuning with Daymet data, with performance comparable to the reference gridded product.

WTD: The ConvLSTM model transferred from Europe reproduced the seasonal dynamics of groundwater fluctuations across the U.S. domain, with anomaly-based evaluation against USGS in-situ wells confirming positive correlations at most monitoring locations.

Groundwater quality: MLMapper 2.0 achieved test accuracy above 0.94 in Chile and between 0.70 and 0.84 in Madagascar, demonstrating the tool's adaptability to different hydrogeological settings and data availability conditions.

Reservoir storage: The ensemble-tree model showed that multi-site training improves forecast skill for extreme events, although direct transfer to reservoirs with different operational characteristics offered limited improvement over a simple benchmark without local retraining.

The results also show that globally available reanalysis forcing products can sustain reasonable machine learning model performance even without high-resolution regional data, though targeted bias correction and fine-tuning with local data remain the most effective ways to improve skill where it matters operationally. Across all four modelling frameworks, the degree of transferability is strongly governed by the hydroclimatic and operational similarity between the training and target domains.

These findings position data-driven and deep learning approaches as flexible, transferable, and physically grounded tools within the STARS4Water modelling framework, capable of bridging basin-scale process understanding and pan-European water resources assessment. They provide a basis for the wider application of machine learning models in other river basins.

Table of contents

Summary	iii
1. Introduction	1
1.1. Objectives of this report	1
1.2. The STARS4Water project	1
1.3. Work Package 5	2
1.4. Content of this report	3
2. Predicting daily SWE in the Rocky Mountains, US.....	5
2.1. Introduction.....	5
2.2. Approach	5
2.3. Results	7
2.4. Conclusions.....	12
3. Estimating daily water table depth in US	14
3.1. Introduction.....	14
3.2. Region and Data	14
3.3. Results	18
3.4. Conclusions.....	23
4. Predictive mapping of groundwater-related variables.....	25
4.1. Introduction.....	25
4.2. MLMapper application in Chile	25
4.3. MLMapper application in Madagascar.....	28
4.4. Conclusions.....	31
5. Ensemble-tree models for reservoir storage	32
5.1. Introduction.....	32
5.2. Region and Data	32
5.3. Method	35
5.4. Results	36
5.5. Conclusions.....	39
6. General conclusions and recommendations for wider uptake.....	40
7. Bibliography	42

1. Introduction

1.1. Objectives of this report

Climate change is altering the timing and magnitude of snow accumulation, groundwater recharge, and surface water availability across large parts of the globe, with direct consequences for water supply, hydropower generation, agriculture, and ecosystems. Reliable high-resolution information on water resources is therefore needed not only at the river basin scale, but also at continental and global scales, where traditional physically based models become computationally prohibitive and observation networks are unevenly distributed. In this context, one of the main objectives of the STARS4Water project is to develop tools based on data-driven models that are not limited to a single river basin but act as general models capable of estimating hydrological variables in other regions and under different climatic conditions. Data-driven and machine learning surrogate models offer a computationally affordable way to achieve this, if models trained over relatively small domains can be transferred to much larger and climatically diverse regions without excessive loss of predictive skill.

In Deliverable D4.4 (Avila et al., 2026a) the upscaling and transfer learning capacity of the STARS4Water machine learning models was assessed at the European scale. ConvLSTM surrogate frameworks trained at the river basin scale, were shown to generalize across Norway, the European Alps, and the pan-European domain, retaining performance levels broadly comparable to the reference physically based models. In parallel, the MLMapper predictive mapping tool demonstrated effective generalisation for groundwater quality assessment across European aquifer systems, and an ensemble-tree model for reservoir storage forecasting was developed and validated for reservoirs in the Duero, East Anglian and Seine river basins. A key question that remained open was whether these transfer capacities extend beyond Europe, to regions characterized by different hydroclimatic regimes, topography, aquifer systems, and reservoir operational practices, and where high-resolution local reference data may be scarce.

The present report addresses this question by extending the analyses carried out in EU river basins by applying and evaluating the four machine learning frameworks to regions outside Europe. The model outputs developed within this study are available through the STARS4Water Zenodo community ([link](#)), which hosts the datasets produced in this project. By making these resources openly accessible, the project aims to support transparency and reproducibility, and to facilitate their reuse and further adaptation by the wider research and water management community beyond the scope of this report.

1.2. The STARS4Water project

The STARS4Water project (Supporting Stakeholders for Adaptive, Resilient, and Sustainable Water Management) is a 4-year initiative funded by the Horizon Europe Program. It focuses on advancing the understanding, observation, and monitoring of water resource availability, contributing to the objectives of the European Green Deal and EU water policies. The project pursues two main goals:

- Enhancing knowledge of how climate change impacts water resources at the river basin level, while identifying vulnerabilities affecting ecosystems, communities, and economic sectors.
- Creating and delivering innovative data services and models to empower decision-makers with tools for planning and implementing adaptive, resilient, and sustainable freshwater management strategies.

Through these efforts, STARS4Water seeks to bridge knowledge gaps and provide actionable insights for sustainable water resource governance. Therefore, the STARS4Water project is built around two core components, as shown in Figure 1.1. The first focuses on fostering collaboration with seven distinct river basin hubs (RBHs): the Danube, Drammen, Duero, East Anglia, Messara, Rhine and Seine (Figure 1.2). Using a living lab-inspired, co-creation approach, the project engages stakeholders in designing tailored services and models. This participatory process ensures that the tools and solutions developed align closely with stakeholder needs, enhancing their utility and longevity beyond the project's timeframe. The second component emphasizes the integration of models into existing tools used for river basin management and decision-making. By leveraging these innovations, the project enables more accurate forecasting of water resource availability and introduces fresh perspectives on the complex relationships between water, nature, and society. These advancements expand the scope of indicators available, supporting more informed and holistic water management strategies.

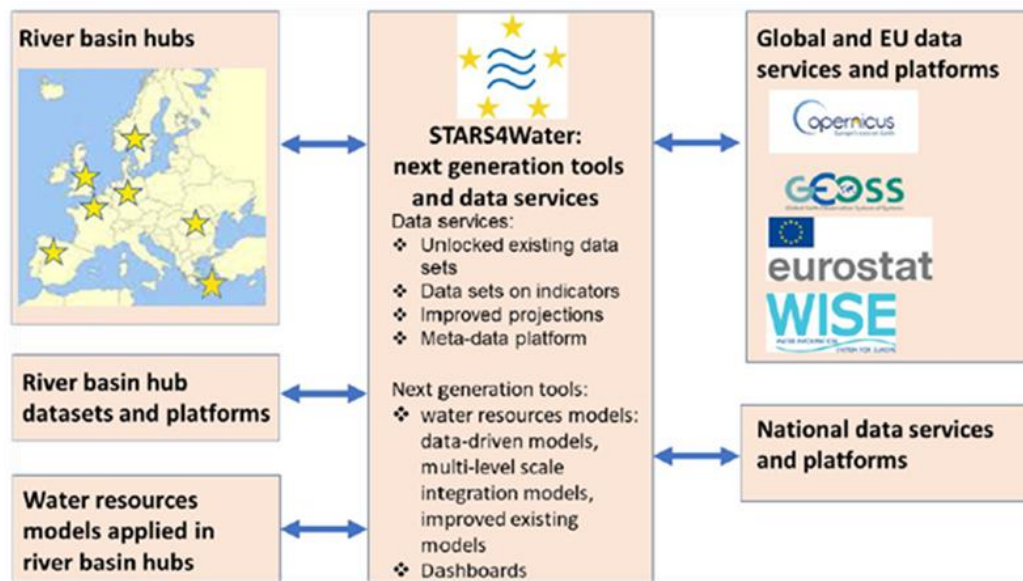


Figure 1.1. Overview of STARS4Water activities within the context of stakeholders and data providers.

1.3. Work Package 5

Work Package 5 (WP5) of the project is dedicated to upscaling and uptake for preparedness. The main objective is to synthesise the results from risk assessments at river basin and pan-European level conducted in Work Package 4 (WP4) and to promote uptake of the project's outcomes by stakeholders and the wider water management community. The main activities in this work package are to:

- Synthesize project results for guidance to practitioners' climate risk assessments (Task 5.1).
- Upscale experiences to other river basins inside and outside the EU, based on identification of similarities in climatic, abiotic, biotic, economic and governance characteristics (Task 5.2).
- Document the advancements in datasets, data-driven models and dashboards to promote sustainable uptake of results in the 7 river basin hubs, generalising findings where applicable to make them relevant for a wider range of basins (Task 5.3).
- Create a storyline map capturing main insights and recommendations for adaptation in co-design with stakeholders, which can be tailored to the specific needs of any river basin in Europe and abroad (Task 5.4).

- Prepare a blueprint containing recommendations for strengthening EU policies and actions based on the results of risk assessment at river basin level and pan-European scale (Task 5.5).



Figure 1.2. Overview of the seven river basin hubs of the STARS4Water project and the main sectors impacted by changes in water resources availability due to changing climate and socio-economic developments.

The interrelation of Work Package 5 in the STARS4Water Framework is displayed in Figure 1.3. Its activities are strongly supported by stakeholders' co-design (WP1) and rely on data (WP2) and models (WP3) developed along the integrated assessment of future water resources carried out in WP4.

This report presents the results of the activities of task 5.2: upscaling machine learning models to other river basins outside Europe. Results of upscaling to EU river basins are reported by Avila et al. (2026a).

1.4. Content of this report

This report extends the application of the machine learning models presented in D.3.3 (Avila & Kollet, 2026), D3.4 (Baron & Keller, 2026) and D4.4 (Avila et al., 2026) by upscaling and transferring the prediction task to regions outside Europe. In particular, the report presents the following analyses:

- ConvLSTM surrogate model for daily Snow Water Equivalent (SWE): transfer from European training domains (Drammen River Basin, Norway) to the Rocky Mountains, US, including fine-tuning with Daymet data and validation against independent in-situ snow observations.
- ConvLSTM surrogate model for daily Water Table Depth (WTD): transfer from Central Europe (ParFlow/CLM DE06) to the contiguous United States using ParFlow-CONUS1 as reference, with evaluation against USGS in-situ well observations and comparison of NLDAS-2 and ERA5 forcing configurations.

D5.3: UPSCALING MACHINE LEARNING MODELS TO OTHER RIVER BASINS

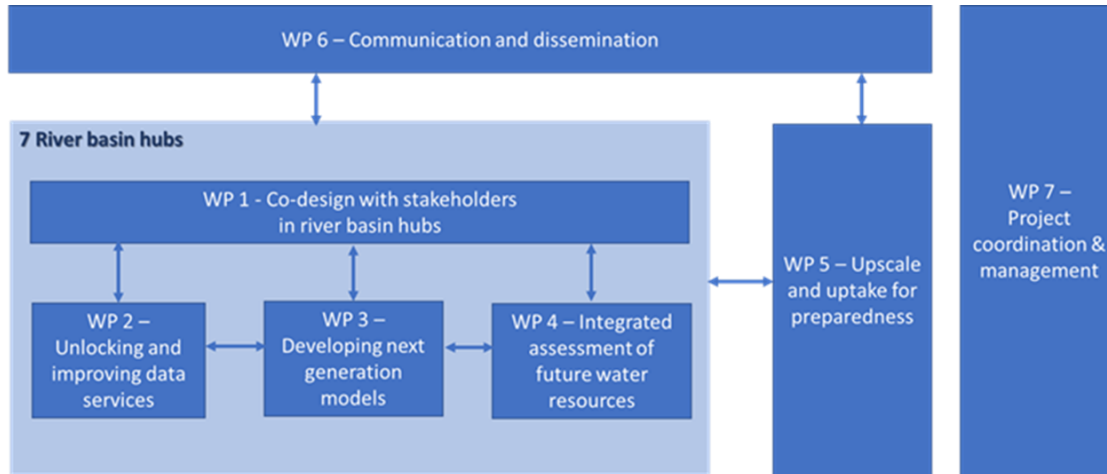


Figure 1.3. Work package structure of the STARS4Water project: the development of the next generation models (WP3) through a co-creation approach (WP1), benefiting from data sources available (WP2) and calibration and validation activities in WP4.

- MLMapper 2.0 predictive mapping tool: application to groundwater-dependent ecosystem identification in the Ligua-Petorca basins, Chile, and groundwater exploitation suitability assessment in southern Madagascar, demonstrating cross-continental generalisation of ensemble tree algorithms for groundwater quality and availability.
- Ensemble-tree model for monthly reservoir storage forecasting: extension from European reservoirs (Duero, UK, Seine basins) to tropical reservoirs in Thailand, evaluating single-reservoir, regional, and cross-continental training configurations through leave-one-out analysis.

The report is structured as follows. Chapter 2 presents the ConvLSTM surrogate model for daily SWE prediction, covering its transfer from European training domains to the Rocky Mountains. Chapter 3 describes the ConvLSTM surrogate for daily WTD estimation transferred to the contiguous United States, including validation against USGS in-situ observations. Chapter 4 presents the MLMapper 2.0 applications to groundwater-dependent ecosystem mapping in Chile and groundwater suitability assessment in Madagascar. Chapter 5 describes the ensemble-tree model for reservoir storage forecasting extended to Thailand. Chapter 6 summarizes the main findings and their implications for the broader STARS4Water framework.

2. Predicting daily SWE in the Rocky Mountains, US

2.1. Introduction

This chapter presents the results obtained from a ConvLSTM to predict daily SWE in the Rocky Mountains. The model development was previously presented in (Avila & Kollet, 2025). The aim of this model was to predict daily Snow Water Equivalent in the Drammen River Basin. The model was later transferred to full Norway and to the European Alps (Avila et al., 2026a). In this study, an initial verification showed that the original pre-trained model was unable to adequately reproduce the daily SWE dynamics of the Rocky Mountains when applied directly. Therefore, unlike previous applications in D4.4, the ConvLSTM framework was applied here using a gradual fine-tuning process, in which selected layers of the network were retrained with Daymet dataset, a high-resolution (1 km) gridded daily climate product for North America (Thornton et al., 2013; Thornton et al., 2016) to adapt the model weights to the snow conditions of this new domain.

2.2. Approach

The Rocky Mountains represent one of the most important snow-dominated hydrological systems in North America, providing critical water resources for downstream communities, agriculture, and ecosystems across the western United States. Seasonal snowpack in this region acts as a natural reservoir, storing precipitation during winter months and releasing it gradually during spring and summer melt periods (Barnett et al., 2005; Musselman et al., 2021). Accurate SWE estimation is therefore essential for water supply forecasting, flood risk management, and long-term water resources planning.

The study area covers the part of the Rocky Mountains domain of the contiguous United States, spanning approximately from 130°W to 90°W and from 30°N to 60°N (Figure 2.1). The region is characterized by complex topography system, with elevations ranging from approximately 500 m in the foothills to over 3,000 m at the highest peaks, producing strong spatial variability in temperature, precipitation, and snow accumulation dynamics, making the domain a demanding test case for evaluating the transferability of the ConvLSTM surrogate model from European training conditions to North American mountain environments.

The hydroclimate of the Rocky Mountains is dominated by orographic enhancement of precipitation, with snowfall constituting a significant proportion of annual precipitation at elevations above approximately 1,500–2,000 m. Annual precipitation varies considerably across the domain, ranging from less than 300 mm in the rain shadow areas east of the continental divide to over 1,500 mm on the windward western slopes. Mean annual temperatures decrease with elevation, and the snow season typically extends from October–November at high elevations to May–June, with some high-altitude sites retaining snow cover year-round. This combination of steep topographic gradients, strong precipitation gradients, and pronounced seasonal snow dynamics makes the Rocky Mountains a demanding and informative test case for evaluating the transferability of the ConvLSTM surrogate model trained in Scandinavian conditions.

The evaluation of the ConvLSTM surrogate model is conducted in two complementary ways. First, model predictions are benchmarked against the Daymet Version 4 daily gridded dataset, which provides spatially continuous estimates of SWE at 1 km resolution over North America. Pixel-wise performance metrics, including Kling–Gupta Efficiency (KGE), Nash–Sutcliffe Efficiency (NSE), Pearson correlation, and the coefficient of determination, are computed to assess how well the surrogate reproduces the spatiotemporal patterns of the reference dataset

across the study domain. In addition, the model predictions are compared against independent in-situ observational records from the NH-SWE dataset (Fontrodona-Bach et al., 2023). Stations within the Rocky Mountains study domain were selected and filtered to ensure spatial overlap with the ConvLSTM prediction grid, providing a rigorous ground-truth test of the model's transferability independent of the gridded reference used for benchmarking.

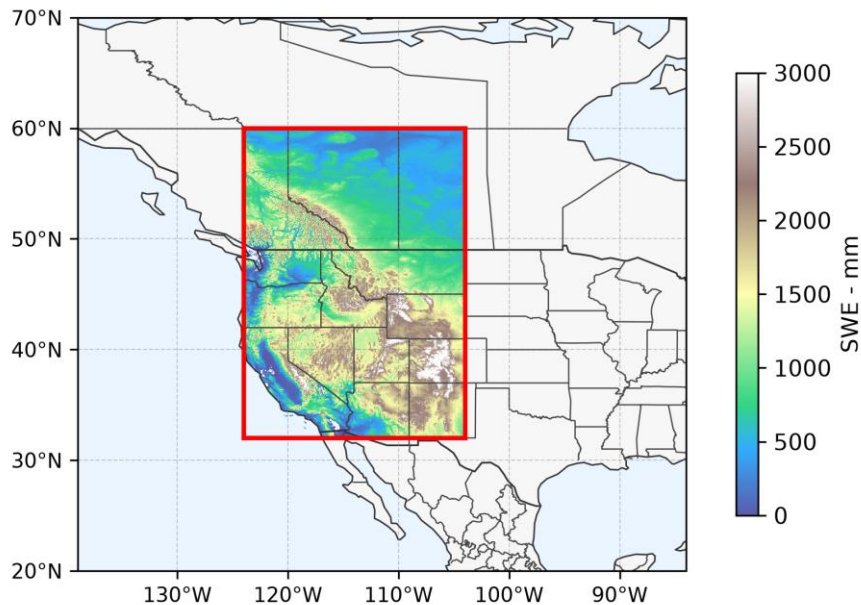


Figure 2.1. Study area used to upscale the ConvLSTM model in US (Rocky Mountains) to predict daily snow water equivalent.

Daymet

Daymet is a gridded dataset of surface meteorological observations produced by the Environmental Sciences Division at Oak Ridge National Laboratory (ORNL). It provides daily estimates of key climate variables over continental North America at $1\text{ km} \times 1\text{ km}$ spatial resolution, covering the period from 1980 to the present. The dataset is constructed by applying a calibrated interpolation scheme to ground-based observations from thousands of weather stations, accounting for the effects of elevation, distance to coastlines, and other physiographic predictors on spatial patterns of temperature and precipitation (Thornton et al., 2013; Thornton et al., 2016). Daymet is freely available through the NASA Earthdata portal (<https://data.ornldaac.earthdata.nasa.gov/>). In this study, Daymet Version 4 Release 1 (V4R1) daily files were downloaded in NetCDF format for the period 2000–2020 and used as the reference benchmark dataset for evaluating the ConvLSTM surrogate model predictions of daily Snow Water Equivalent across the Rocky Mountains.

2.3. Results

Figure 2.2 presents the spatial performance of the ConvLSTM surrogate model across the Rocky Mountains study domain, computed pixel-wise against the Daymet reference dataset. The model was evaluated for the period 2001-2009. For better visualization, the figure presents the performance only in pixels with average SWE greater than 10 mm, since pixels with lower average SWE experience only marginal, intermittent snow cover, where performance metrics such as KGE and NSE become numerically unstable and difficult to interpret due to the near-zero variance in the target signal. The four complementary metrics: Kling–Gupta Efficiency (KGE), Nash–Sutcliffe Efficiency (NSE), Pearson correlation (r), and coefficient of determination (R^2), provide a comprehensive assessment of model skill in reproducing the spatiotemporal patterns of daily SWE. The spatial maps reveal that the ConvLSTM model achieves generally good performance across most of the study domain, with KGE values exceeding 0.6 in most low- to mid-elevation areas where seasonal snow dynamics are well defined.

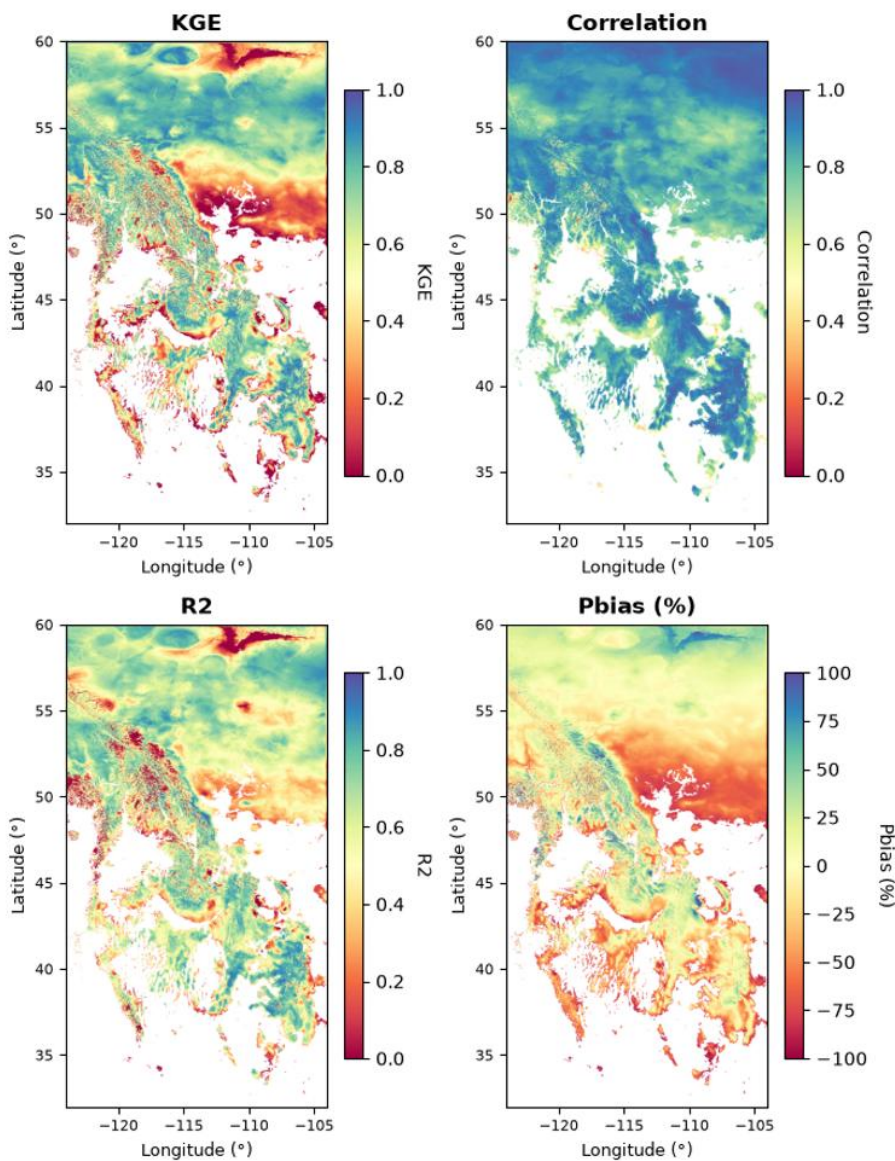


Figure 2.2. Spatial distribution of ConvLSTM model performance metrics evaluated against the Daymet reference dataset across the Rocky Mountains.

Performance metrics are generally higher in areas with moderate topographic complexity, where the ConvLSTM model can effectively capture the elevation-dependent gradients in temperature and precipitation that drive snow accumulation and melt. In contrast, model performance tends to decrease in the highest elevation zones and in areas with complex terrain features such as steep slopes, where local-scale processes such as wind redistribution and radiation-driven melt become more important. These patterns are consistent with the challenges reported in previous studies applying deep learning models to SWE prediction in mountainous environments (Duan et al., 2024; El Halabi et al., 2025).

In addition, Figure 2.3. presents the temporal evaluation daily SWE produced by the ConvLSTM model against the Daymet for the period 2001–2009. The upper panel shows the domain-aggregated daily SWE time series, while the lower panel displays the SWE time series at a representative grid point selected to illustrate local-scale performance. For both cases, the ConvLSTM model reproduces the seasonal timing of snow accumulation and melting, correctly tracking the onset of snow accumulation in autumn, the peak SWE in spring, and the subsequent snowmelt recession through early summer.

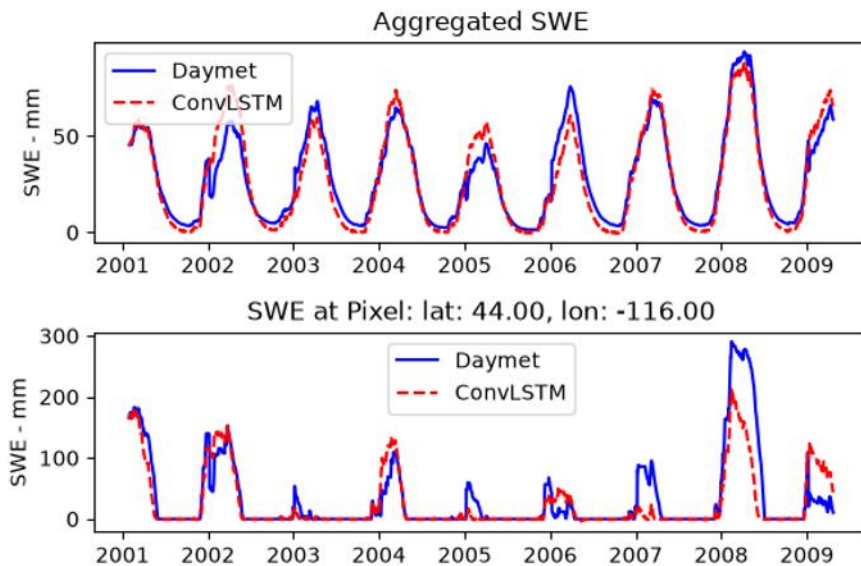


Figure 2.3. Temporal evaluation of the ConvLSTM model against the Daymet reference dataset over the Rocky Mountains. The upper panel shows the domain-aggregated daily SWE time series, and the lower panel presents the SWE time series at a representative grid point.

The domain-aggregated comparison reveals that the ConvLSTM model captures the interannual variability in peak SWE magnitude, although some systematic biases are observed during extreme snow years, particularly during periods of exceptionally high or low accumulation. At individual grid points, the model shows some underestimation compared to the target dataset, reflecting some challenges of reproducing local-scale SWE dynamics in areas with complex topography where elevation-dependent gradients in temperature and precipitation exert a dominant control on snowpack development. However, the temporal evaluation demonstrates that the ConvLSTM model can well generalize its learned snow dynamics from the Norwegian training domain to this North American mountain environment, successfully reproducing the dominant seasonal and interannual patterns of SWE evolution.

Figure 2.4 summarizes the overall performance of the ConvLSTM through boxplots of five key evaluation metrics: NSE, KGE, Pearson correlation (r), variability ratio (α), and bias ratio (β). The model achieves moderate to good results across most of the domain, with median KGE values greater than 0.6, indicating that the ConvLSTM framework can reproduce the main features of daily SWE dynamics. The Pearson correlation component of KGE is generally strong across the domain, indicating that the model effectively captures the temporal dynamics of snow accumulation and melt. The variability ratio (α) shows that the model tends to slightly underestimate the amplitude of SWE fluctuations, which may be related to differences in snow process characteristics between the Norwegian training domain and the Rocky Mountains. The bias ratio (β) also reveals a tendency toward slight underestimation of the mean SWE in some areas, possibly reflecting the influence of different precipitation regimes and snowpack evolution dynamics between the two regions. Finally, the NSE values are generally positive with median values greater than 0.6. However, lower NSE values are observed in the highest elevation zones where snow cover is more persistent and in rain-shadow areas where snow accumulation is more episodic and harder to predict. These performance patterns highlight the importance of local topographic and climatic controls on model skill and suggest that targeted fine-tuning with local data could further improve performance in challenging terrain.

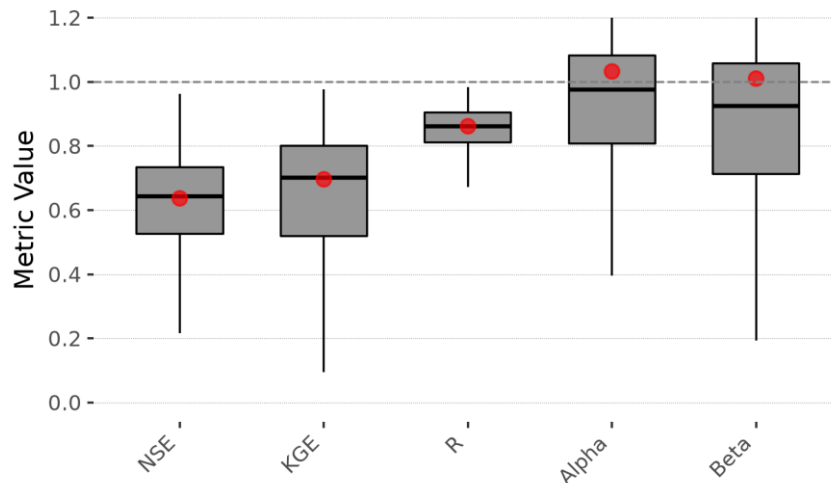


Figure 2.4. Summary of ConvLSTM model performance across the Rocky Mountains evaluated against the Daymet reference dataset. Boxplots show the distribution of NSE, KGE, Pearson correlation (r), variability ratio (α), and bias ratio (β). Red markers indicate median.

In addition to the evaluation against the Daymet gridded reference dataset, the ConvLSTM model performance was assessed against independent observational records to provide a more rigorous test of model transferability. Figure 2.5 presents a spatial comparison of KGE and Figure 2.6 presents the Pearson correlation coefficients evaluated against available in-situ snow observations.

The comparison reveals that the ConvLSTM model achieves performance levels that are broadly consistent with those of the Daymet reference dataset at observation stations, suggesting that the surrogate framework has successfully learned the dominant controls on SWE variability from the training data. In some areas, the ConvLSTM model shows improved performance relative to Daymet, particularly in capturing peak SWE timing and magnitude, which may reflect the model's ability to learn nonlinear relationships between meteorological forcing and snowpack response that are not fully captured by the statistical interpolation methods underlying Daymet.

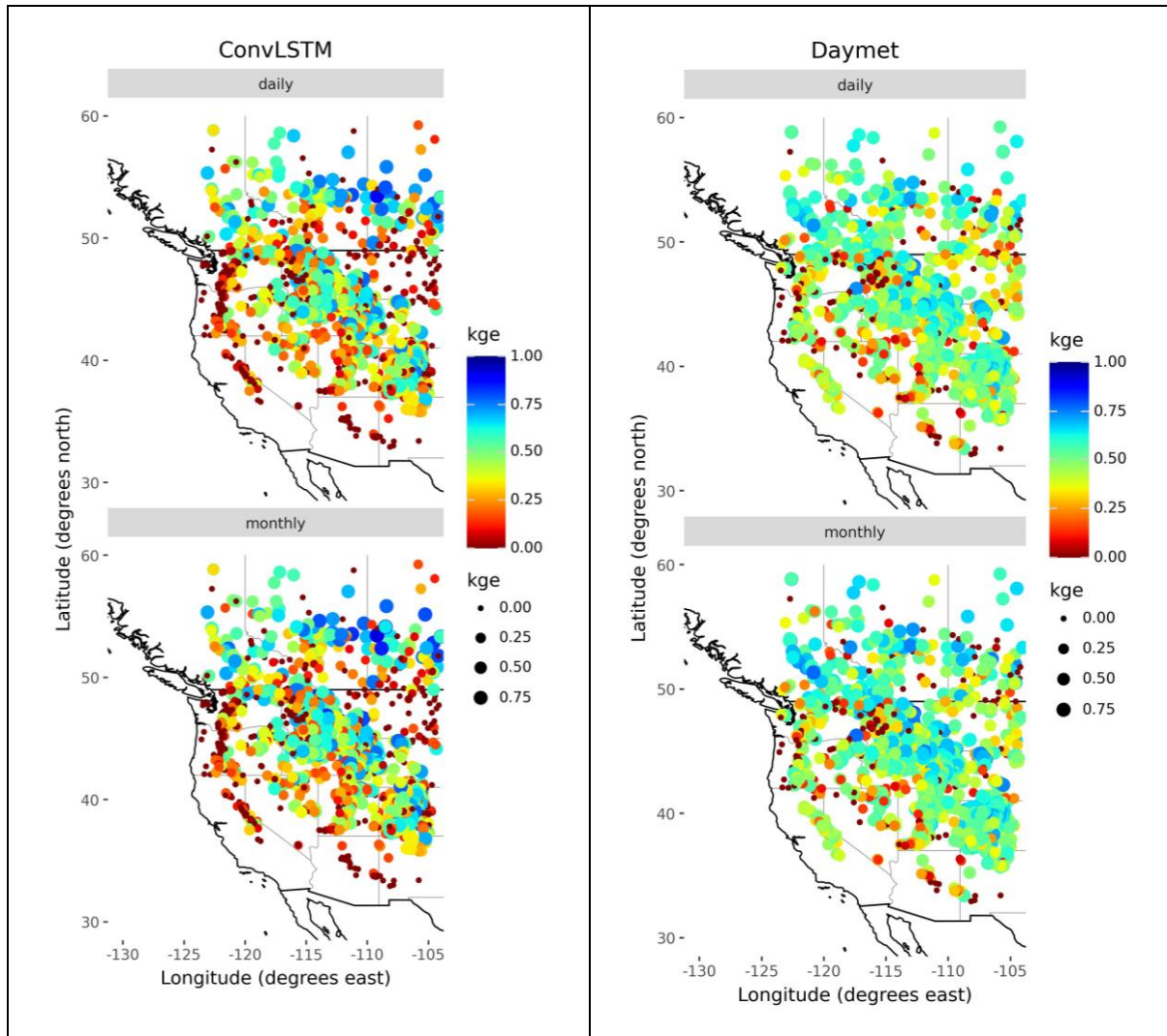


Figure 2.5. KGE comparison between simulated SWE derived from ConvLSTM and Daymet, evaluated against independent observational data in the Rocky Mountains.

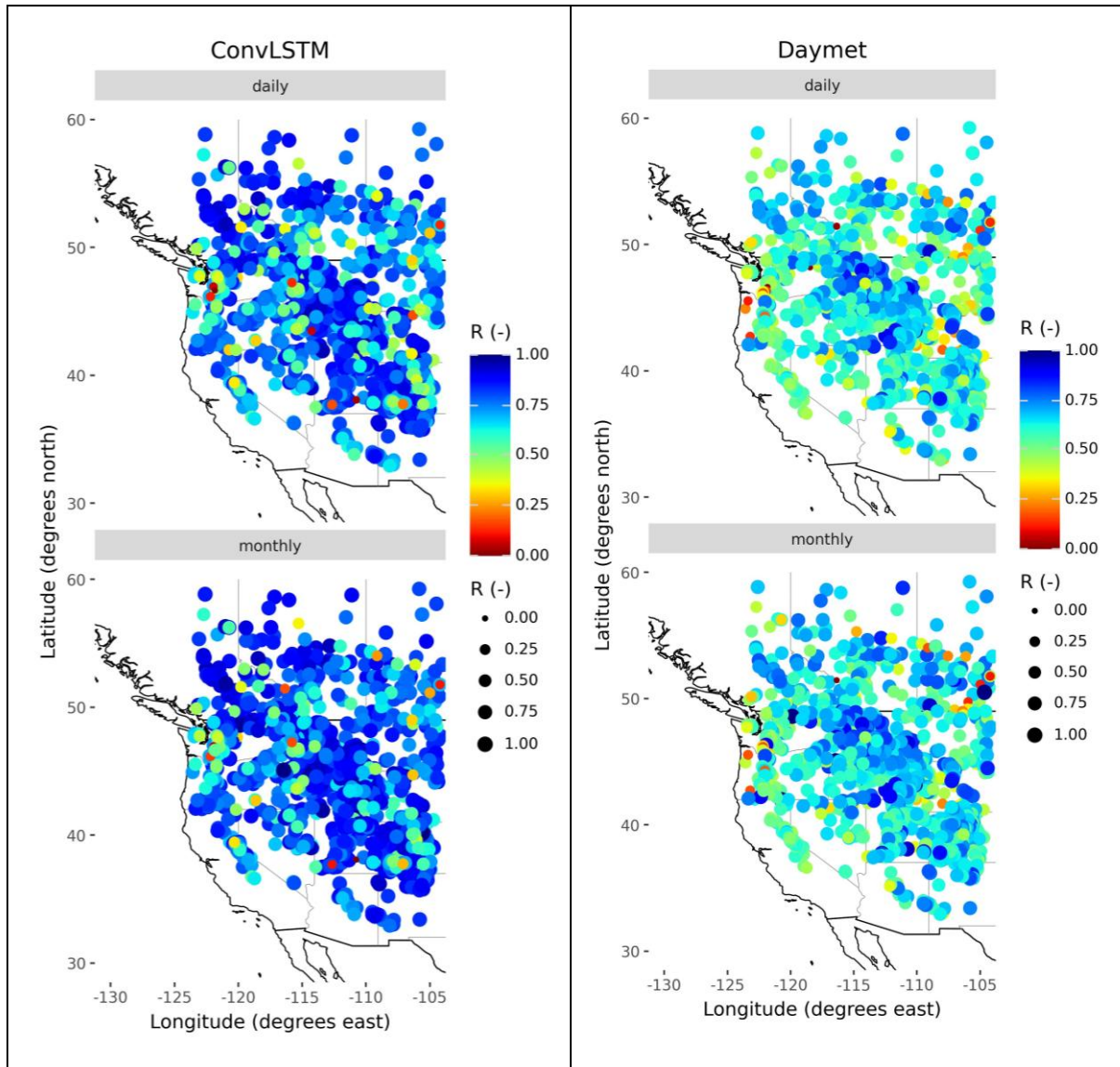


Figure 2.6. Pearson correlation comparison between simulated SWE derived from ConvLSTM and Daymet, evaluated against independent observational data in the Rocky Mountains.

Figure 2.7 provides a quantitative synthesis of model performance against observational data through boxplots of KGE and its three decomposed components (correlation, variability ratio, and bias ratio). The results indicate that the ConvLSTM model achieves reasonable agreement with observations, with median KGE values that demonstrate the practical utility of the transferred surrogate model for SWE estimation in the Rocky Mountains.

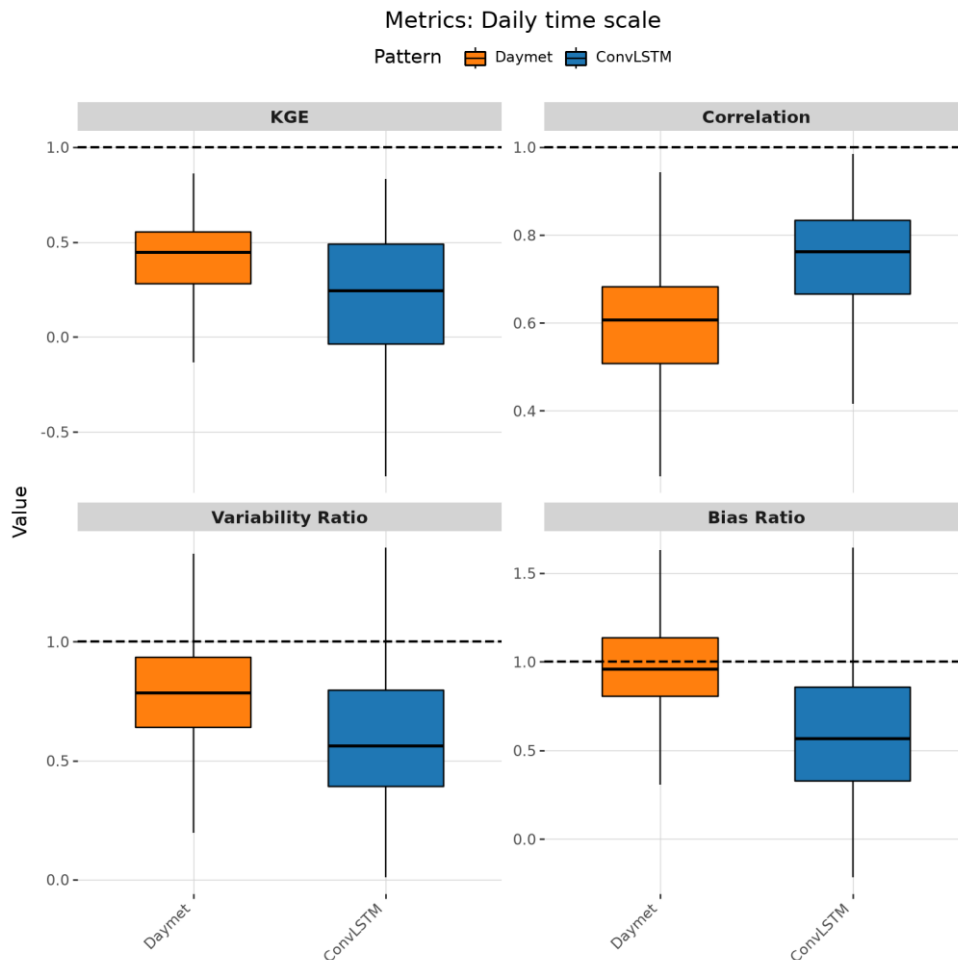


Figure 2.7. Summary performance comparison between simulated SWE derived from ConvLSTM and Daymet, evaluated against observational data in the Rocky Mountains. Boxplots show KGE and its three decomposed components (correlation, variability ratio, and bias ratio).

2.4. Conclusions

The overall performance of the ConvLSTM model transferred to the Rocky Mountains confirms the viability of the surrogate modelling approach for cross-continental SWE prediction. When fine-tuned with Daymet reference data, the model achieved median KGE values above 0.6 across most of the study domain, with the highest skill concentrated in low to mid-elevation areas characterized by well-defined seasonal snow dynamics. These results are comparable to, though somewhat lower than, the performance reported for European domains, where median KGE exceeded 0.75 in the Drammen River Basin (Avila et al., 2026b). The reduction in skill is expected given the substantially greater climatic and topographic distance between the Norwegian training domain and the Rocky Mountains.

A key advantage of the ConvLSTM framework is its ability to capture spatiotemporal dependencies in SWE evolution through recurrent convolutional layers, which enables it to learn patterns of snow accumulation and melt that remain largely physically consistent within the observed range of conditions, without requiring explicit process-based formulations. This data-driven approach also offers significant computational efficiency compared to physically based snow models, enabling rapid SWE predictions across large domains at daily temporal resolution.

Despite these encouraging results, several limitations of the current implementation should be acknowledged. First, the need for fine-tuning represents a practical constraint: the pre-trained Drammen model was unable to adequately reproduce SWE dynamics in the Rocky Mountains when applied directly, exhibiting significant biases in peak magnitude and melt timing. This indicates that the model's transferability is not unconditional and depends on the degree of hydroclimatic similarity between training and target domains. Second, model performance degrades in the highest elevation zones and areas of complex terrain, where Daymet reference dataset may not fully capture the fine-scale variability in temperature and precipitation that drives snowpack development. Third, ConvLSTM architecture does not explicitly incorporate physical constraints such as energy balance or mass conservation, which limits its interpretability and may lead to physically inconsistent predictions under extreme or novel climatic conditions. Fourth, the evaluation is performed against the same Daymet reference dataset used for the fine-tuning, which may yield optimistic skill estimates; validation against fully independent observations or alternative reference datasets therefore remains necessary.

Several avenues for future improvement can be identified. The integration of additional static features such as slope, aspect, and land cover type as auxiliary inputs to the ConvLSTM architecture may enhance performance in topographically complex areas by providing the model with explicit information about terrain controls on snow distribution. Exploring alternative fine-tuning strategies, including progressive freezing of early convolutional layers and domain-adversarial training, could improve transfer efficiency and reduce the amount of target-domain data required for adaptation. The use of other reference datasets, such as SNODAS (Barret, A.P, 2003) or snow reanalysis products, would allow a more rigorous evaluation of model performance at fine spatial scales. Furthermore, extending the assessment to include other mountain ranges globally, such as the Andes and the Himalayas, would provide a more comprehensive understanding of the model's generalization capacity across different snow climates. Finally, hybrid approaches that combine ConvLSTM predictions with physically based snow models through data assimilation frameworks could leverage the strengths of both paradigms, improve accuracy while maintain computational efficiency. These developments would further establish surrogate modelling as a practical tool for operational SWE monitoring and water resources management in snow-dominated regions worldwide.

3. Estimating daily water table depth in US

3.1. Introduction

This chapter extends the transferability assessment of deep learning surrogate models for Water Table Depth (WTD) beyond the European domain to the United States. In Avila et al. (2026a), a ConvLSTM-based machine learning model framework was trained on high-resolution ParFlowCLM DE06 simulations over Central Europe, forced with the ECMWF High-Resolution (HRES) atmospheric reanalysis, and subsequently transferred, without retraining, to the full pan-European domain using TSMP simulations as an independent benchmark. The results showed that the model retained strong predictive skill across markedly different spatial scales and hydrogeological settings, achieving median NSE and KGE values near 0.7 and Pearson correlation exceeding 0.8 in some regions across Europe, with the strongest performance concentrated in flat, recharge-dominated lowlands, and a systematic degradation in topographically complex regions such as the Alps and the Mediterranean.

Here, we aim to assess whether the same transfer learning capacity holds when the target domain shifts from a continental European setting to a hydro-climatically distinct region such as the United States, characterized by markedly different topography, aquifer systems, and climate zones. To this end, the present study applies the previously developed ConvLSTM framework to the contiguous United States, using WTD and associated hydrologic states simulated by the ParFlow-CONUS1 model (Maxwell et al., 2015; O'Neill et al., 2021) as the reference dataset. ParFlow-CONUS1 provides a physically consistent, high-resolution (1 km) integrated representation of coupled groundwater–surface water–land–atmosphere processes across the entire U.S. domain, evaluated against thousands of in-situ observations and remote sensing products, making it a robust benchmark for testing the cross-continental generalization of the ConvLSTM surrogate. To assess how sensitive this transferability is to the meteorological forcing used, the ConvLSTM model was applied in two configurations: one forced with NLDAS-2, a regionally calibrated, high-resolution reanalysis product widely used for hydrologic modeling over the U.S., and another forced with the ERA5-Land global reanalysis, which offers coarser but globally consistent coverage. Comparing these two configurations allows us to isolate the extent to which model performance depends on the availability of high-quality regional forcing versus a globally available, lower-resolution alternative.

By evaluating whether a model trained on European hydrometeorological patterns can reproduce ParFlow-CONUS1-simulated WTD dynamics in the United States without retraining, this study seeks to determine the broader applicability of data-driven transfer learning tools for groundwater monitoring at the global scale.

3.2. Region and Data

The selected area for this study corresponds to a subdomain of the ParFlow-CONUS1 grid, showed in Figure 3.1, and covering approximately the region between 105–95°W and 37–43°N. As shown by the simulated WTD field, the eastern and northern portions of the domain are dominated by shallow, precipitation-responsive water tables (purple tones), whereas pronounced deep-WTD streaks (>20–30 m) appear concentrated in the central and western portions of the domain.

This hydroclimatic and topographic gradient is further illustrated in Figure . Average daily precipitation decreases sharply from east to west, consistent with the orographic influence of the Rocky Mountain foothills, while topography indicates the transition from the low-relief Great Plains in the east to elevations exceeding 3000 m in the west.

The volumetric water content anomaly reveals additional spatial heterogeneity within the domain, with wetter anomalies concentrated toward the eastern edge and more variable, patchy conditions in the central and western portions, mirroring the contrasts already evident in the WTD field. This combination of a strong anthropogenic groundwater signal, sharp spatial heterogeneity, and hydroclimatic contrast make the domain a particularly demanding test case within the broader U.S. assessment, placing considerable stress on the surrogate's ability to reproduce local-scale groundwater dynamics under conditions not represented in the training data.

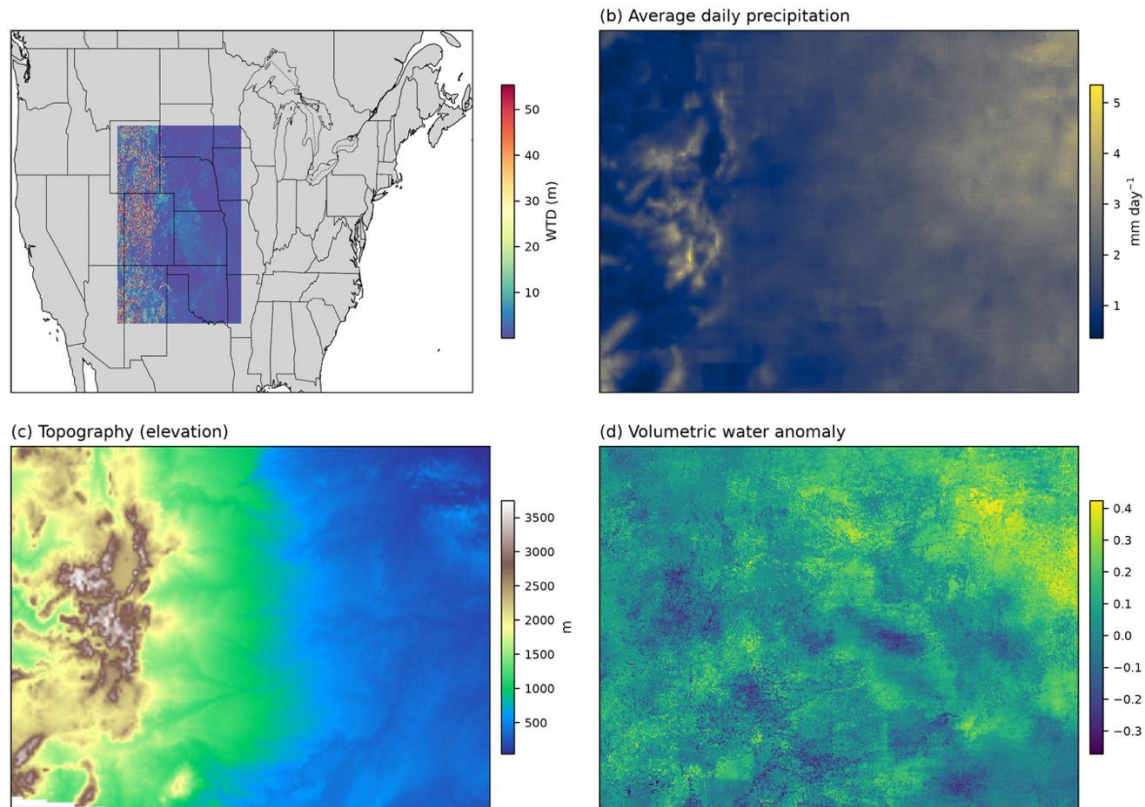


Figure 3.1. Study area for the ConvLSTM WTD transferability assessment. (a) Simulated water table depth field from ParFlow-CONUS1 over the U.S. study domain (105–95°W, 37–43°N); (b) average daily precipitation; (c) topographic elevation; and (d) volumetric water anomalies.

Figure 3.2 presents the daily climatology (2003–2005) of water table depth (WTD), volumetric water content (VWC) anomaly, and precipitation across the study domain. For each variable, the solid line shows the domain-mean daily value, and the shaded band shows the corresponding min–max range across the domain, illustrating both the mean seasonal cycle and its spatial spread. The three variables exhibit a coherent seasonal cycle driven by the domain's precipitation regime. Mean precipitation increases markedly from spring through summer (May–September), with individual events within the domain reaching up to approximately 13 mm/day, and the min–max range widening accordingly as the spatial contrast between wetter and drier areas intensifies. This precipitation increase is accompanied by a corresponding rise in the mean VWC anomaly and a progressive shallowing of the mean water table, from approximately 5.2 m in January to a seasonal minimum of approximately 4.75 m around July; the WTD min–max range widens substantially over this period (reaching roughly 4.2–5.6 m), reflecting pronounced spatial heterogeneity in the timing and magnitude of the recharge response across the domain.

From late summer into autumn and winter, precipitation becomes less frequent and less intense, the mean VWC anomaly turns negative, and the mean water table deepens again, returning to approximately 5.0–5.1 m by December, while the WTD range narrows relative to the summer peak. This recharge–depletion cycle is consistent with a shallow, precipitation-responsive water table regime, and provides a baseline temporal signal against which the performance of the ConvLSTM surrogate model is evaluated in the subsequent sections.

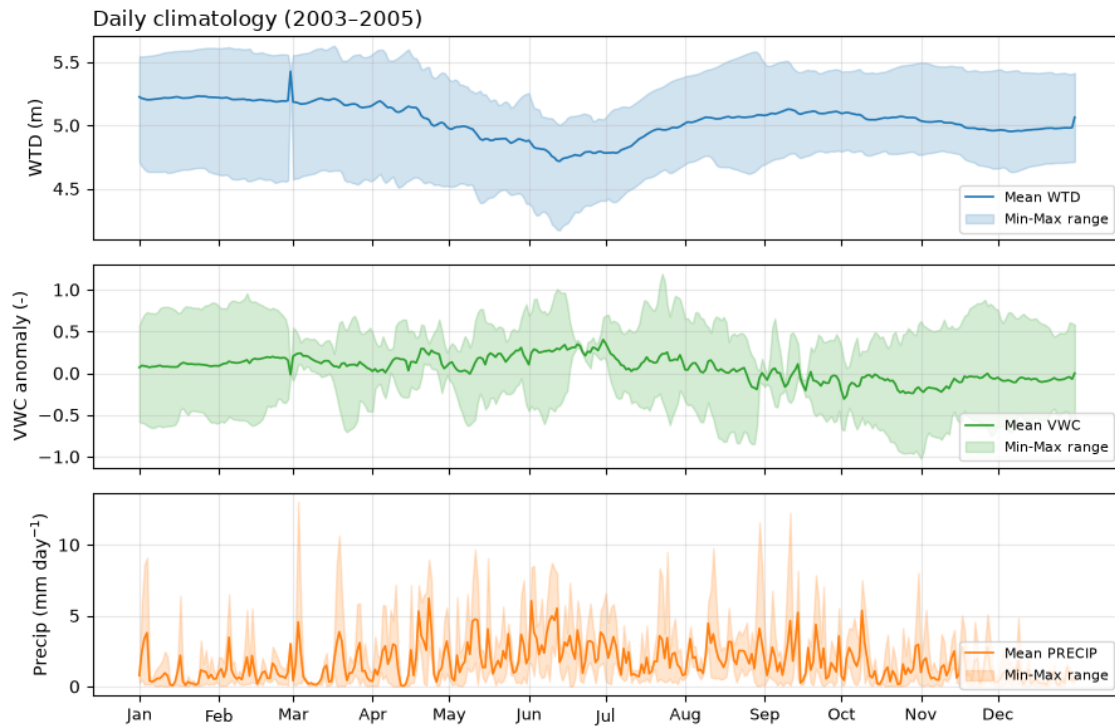


Figure 3.2. Daily climatology (2003–2005) of water table depth (WTD), volumetric water content (VWC) anomaly, and precipitation across the U.S. study domain. Solid lines show domain-mean values; shaded areas indicate spatial variability across the domain.

ParFlow-CONUS1

ParFlow-CONUS1 is a physically based integrated hydrologic model that couples the three-dimensional variably saturated groundwater flow model ParFlow with the Common Land Model (CLM) to jointly simulate the coupled water and energy budgets across the contiguous United States (Maxwell and Miller, 2005; Kollet and Maxwell, 2006). The model solves the mixed form of the Richards equation to represent three-dimensional subsurface flow, using a free-surface overland-flow boundary condition and the kinematic wave approximation to route surface runoff, while CLM computes the surface energy balance and partitions evapotranspiration into direct soil evaporation and vegetation transpiration based on atmospheric forcing (precipitation, temperature, humidity, wind speed and radiation) (O'Neill et al., 2021).

The ParFlow-CONUS1 domain spans approximately 6.3 million km² at 1 km lateral grid resolution (3442 × 1888 cells), with five vertical layers reaching a cumulative depth of 102 m and encompasses the majority of the major U.S. river basins (Maxwell et al., 2015; O'Neill et al., 2021). Model inputs include topographic slopes derived from the HydroSHEDS digital elevation model (Lehner et al., 2008), vegetation classes from IGBP/USGS land cover products, soil hydraulic parameters from the Soil

Survey Geographic database, and subsurface permeability derived from national-scale geologic maps (Gleeson et al., 2011). The model is initialized through an intensive spin-up procedure combining a steady-state groundwater run with a multi-year transient simulation under climatologically average atmospheric forcing, and it is typically driven at hourly resolution using bias-adjusted NLDAS-2 meteorological data blended with higher-resolution radar and satellite products (Cosgrove, 2003; Xia et al., 2012, 2015).

ParFlow-CONUS1 does not explicitly represent anthropogenic processes such as groundwater pumping, irrigation or reservoir operation, but it provides a physically consistent, high-resolution representation of the coupled groundwater–surface water–land–atmosphere system that has been validated against streamflow, evapotranspiration, soil moisture, snow water equivalent, water table depth and total water storage observations, making it a suitable reference dataset for benchmarking and training data-driven models, such as ConvLSTM, outside the European river basin hubs of the STARS4Water project.

HydroData

The ParFlow-CONUS1 simulation outputs and associated input datasets used in this study were obtained from HydroData (Defnet et al., 2024). Through the `hf_hydrodata` Python package and its associated API, HydroFrame provides unified, open access to hydrologic simulations generated from the ParFlow-CONUS1 and ParFlow-CONUS2, together with a wide range of gridded model input datasets (e.g., topography, soil and geologic parameters, land cover, and atmospheric forcing such as NLDAS-2) and point-scale observations from networks such as USGS streamflow gauges, SNOTEL, and AmeriFlux. For the purposes of this study, ParFlow-CONUS1 outputs and forcing variables were retrieved through the HydroData catalog to serve as the reference hydrologic dataset for training and evaluating the ConvLSTM data-driven model. It is worth noting that given some limitations with the HydroData portal, the ParFlow-CONUS1 data, including input forcing and in-situ station data, were obtained from late 2002 to mid-2006, which represents a limitation when comparing the model, since this relatively short period constrains the evaluation of groundwater temporal dynamics in certain regions.

NLDAS

NLDAS (Xia et al., 2012) is a set of reanalysis-based, gridded meteorological forcing products used to drive land-surface and hydrologic models across North America. NLDAS-2 was used to force the ParFlow-CONUS1, providing hourly fields of precipitation, air temperature, humidity, wind speed, and shortwave/longwave radiation at roughly 12 km spatial resolution. In this case, we estimated the daily average for all input variables. In the ParFlow-CONUS1 configuration, the base 12 km NLDAS-2 fields were interpolated onto the 1 km model grid. In our case, we used precipitation daily fields derived from NLDAS to force the ConvLSTM. This methodological choice is directly aligned with the transfer-learning strategy developed in D4.4, where the ConvLSTM surrogate was trained with the native forcing of the physically based reference (HRES for the ParFlow/CLM DE06 training domain) and subsequently evaluated with an alternative, coarser global reanalysis. In the present application, NLDAS-2 constitutes the native forcing of the ParFlow-CONUS1 reference, so the NLDAS configuration of the surrogate operates under the same meteorological conditions under which the reference WTD fields were generated.

3.3. Results

Figure 3.3 presents the spatially distributed evaluation of daily Water Table Depth (WTD) predicted by the ConvLSTM surrogate model performance across the U.S. study domain, computed pixel-wise against the ParFlow-CONUS1 reference dataset for the configuration forced with NLDAS-2 meteorological data. The four metrics: Kling–Gupta Efficiency (KGE), Pearson correlation, Root Mean Square Error (RMSE, in meters), and Percentage Bias (Pbias), reveal a heterogeneous but generally skillful model behavior following the direct cross-continental transfer of the European-trained framework.

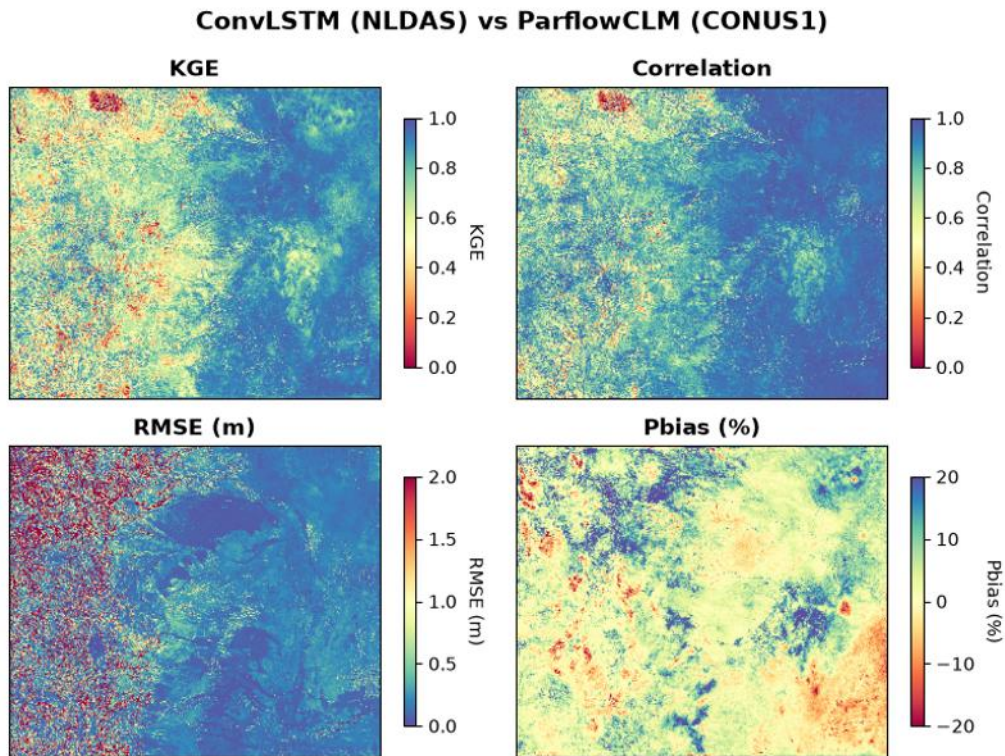


Figure 3.3. Spatial distribution of ConvLSTM surrogate model performance metrics in predicting Water Table Depth (WTD), forced with the NLDAS-2 dataset and evaluated against the ParFlow-CONUS1 reference dataset across the U.S. study domain.

In general, greater KGE values are presented in the eastern and northern portions of the domain, corresponding to the low-relief Great Plains, where shallow, precipitation-responsive water tables and the strong coupling between antecedent precipitation and groundwater dynamics encoded by the Antecedent Precipitation Index closely mirror the recharge-dominated regime of the Central European training domain.

Toward the west, performance degrades progressively, with lower KGE and correlation values and higher RMSE concentrated across the central and western sectors. This western degradation is consistent with the well-documented limitations of the ParFlow-CONUS1 reference itself, whose water table depth evaluation against USGS wells showed the largest “hot spots” of bias in the southern Ogallala aquifer and lower Colorado basin (O'Neill et al., 2021), implying that part of the surrogate error over these deep-water-table regions reflects reference uncertainty rather than surrogate skill alone.

The correlation map also confirms that the temporal phasing of groundwater fluctuations is reproduced across most of the domain, consistent with the

recharge–depletion cycle documented in the climatological in Figure 3.2, while the Pbias map reveals a mixed pattern of positive and negative departures within the $\pm 20\%$ range, with a tendency toward overestimation in the western and central portions of the domain.

Figure 3.4 also presents the corresponding spatial evaluation for the ConvLSTM configuration forced with the ERA5-Land reanalysis dataset. In this case, KGE values are notably lower, particularly across the western region of the study area. The correlation map exhibits a similar, though less pronounced, deterioration, with values below those of the NLDAS-2 configuration spreading across the central and western sectors, indicating that the temporal phasing of groundwater fluctuations is more sensitive to forcing resolution than is the mean state.

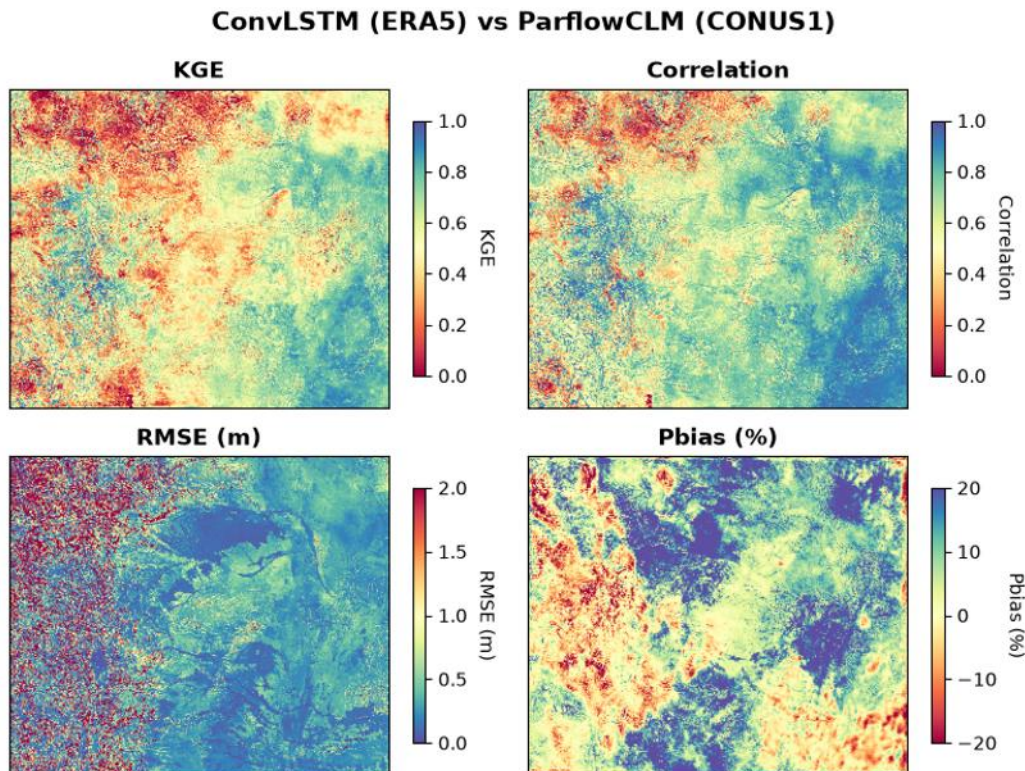


Figure 3.4. Spatial distribution of ConvLSTM surrogate model performance metrics in predicting Water Table Depth (WTD), forced with the ERA5-Land dataset and evaluated against the ParFlow-CONUS1 reference dataset across the U.S. study domain.

The RMSE map reveals errors reaching approximately 2 m in water table depth across the central and western portions of the domain, roughly double the typical errors of the NLDAS-2 configuration in those areas, while the PBIAS map shows broad positive and negative departures that amplify spatial heterogeneity. Notably, whereas the lowest absolute skill occurs in the west, the relative loss of skill introduced by the ERA5-Land forcing is largest precisely where the NLDAS-2 configuration performed best, namely over the shallow, precipitation-responsive aquifers of the eastern Great Plains, suggesting that the high temporal resolution and regional consistency of NLDAS-2 are essential for reproducing the rapid, precipitation-responsive water-table dynamics of the shallow aquifers in that region.

Despite this reduction in skill, the ERA5-Land-forced model retains useful predictive information over substantial portions of the domain and, since it relies exclusively on globally available reanalysis data, demonstrates that the framework can operate even in regions where high-resolution regional forcing products such as NLDAS-2 are unavailable.

Figure 3.5 and Figure 3.6 complement the spatial evaluation with a temporal assessment of both forcing configurations over the period 2003–2005, respectively. For both figures, the upper panels show the domain-aggregated daily WTD time series together with the corresponding anomalies relative to the seasonal climatology, while the lower panels present the WTD time series at a single representative pixel characterized by shallow water table conditions.

At the aggregated scale, both the NLDAS and ERA5-forced configurations reproduce the seasonal cycle of water table depth, tracking the progressive shallowing of the mean water table from approximately 5.2 m in winter to its seasonal minimum of roughly 4.7–4.8 m in late summer and the subsequent deepening toward the end of the year. The anomaly curves reveal that the NLDAS-forced model captures the sign and approximate magnitude of departures from climatology for most of the record, closely following the ParFlow-CONUS1 reference, whereas the ERA5-forced model tracks the same overall phasing but exhibits larger anomaly differences. At the pixel scale, both configurations reproduce the general seasonal behavior and the sign of interannual anomalies, although the ERA5-forced model shows amplified variability and occasional overestimation of peak values during wet periods, while the NLDAS-forced model adheres more closely to the reference simulation. This amplification of variability under coarser forcing reproduces the behavior documented in D4.4 for the European domains, where ERA5-Land-driven surrogates exhibited amplified WTD variability and occasional overestimation of peak values during wet seasons relative to the HRES-forced configuration.

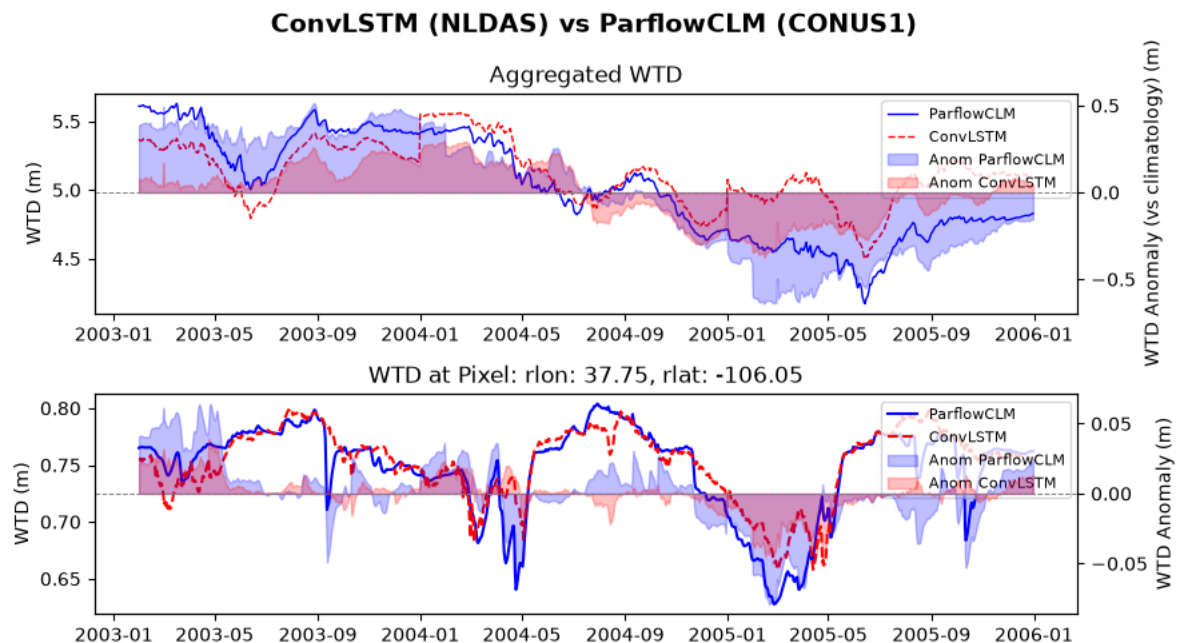


Figure 3.5. Temporal evaluation of the ConvLSTM surrogate model driven with NLDAS-2 forcing data over the period 2003–2005. Upper panels: domain-aggregated daily WTD time series and anomalies relative to seasonal climatology. Lower panels: pixel-scale WTD time series at a representative location within the study domain.

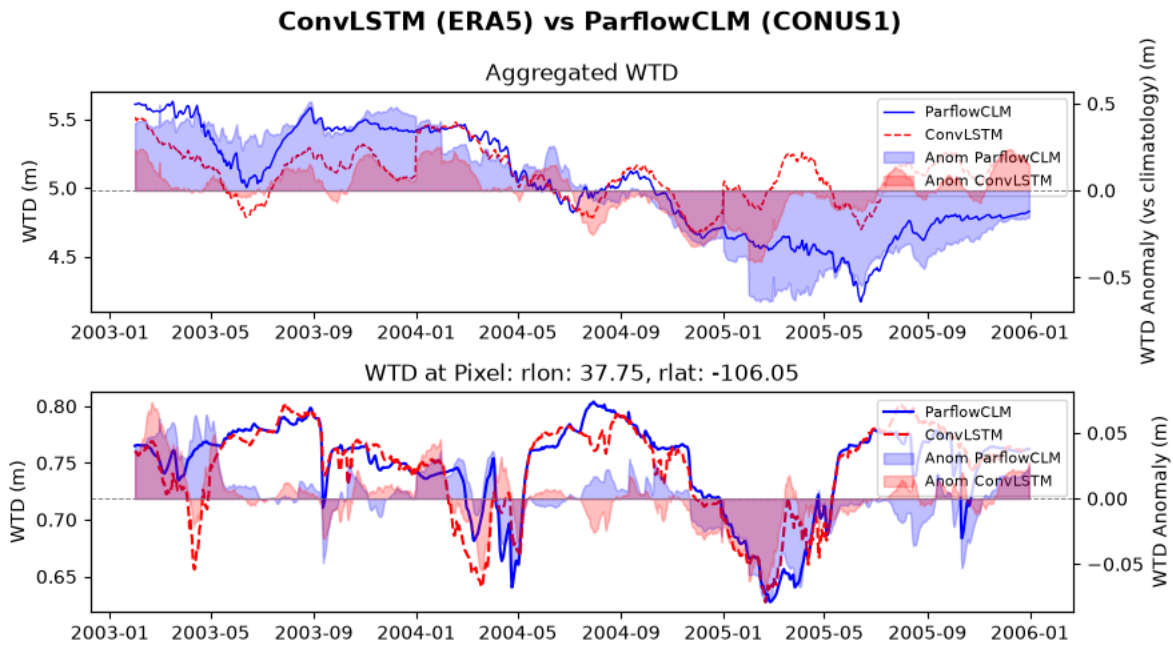


Figure 3.6. Temporal evaluation of the ConvLSTM surrogate model driven with ERA5-Land forcing data the period 2003–2005. Upper panels: domain-aggregated daily WTD time series and anomalies relative to seasonal climatology. Lower panels: pixel-scale WTD time series at a representative location within the study domain.

Figure 3.7 provides a quantitative synthesis of the results through boxplots of the main performance metrics: KGE, NSE, Pearson correlation, coefficient of determination (R^2), variability ratio (α), and bias ratio (β), computed pixel-wise against the ParFlow-CONUS1 reference for the two forcing configurations. A direct comparison with the temporal behavior in Figure and Figure confirms that the differences presented in the time series are systematically reflected in the metric distributions: the NLDAS-forced configuration attains higher median values of KGE, NSE, correlation, and R^2 , with comparatively compact interquartile ranges, whereas the ERA5-forced configuration shows reduced medians and considerably wider spread, mirroring the larger anomaly excursions and occasional offsets observed in the ERA5-Land traces. The variability ratio (α) of the ERA5-Land configuration deviates more strongly from the ideal value of 1.0, consistent with the amplified variability seen at the pixel scale in Figure 3.6, while the bias ratio (β) indicates a systematic bias in mean WTD magnitude for both configurations, with the NLDAS configuration lying closer to zero.

Taken together, the results demonstrate that the ConvLSTM surrogate, transferred from Europe without retraining, reproduces both the timing and the magnitude of groundwater dynamics across the U.S. domain. The results also confirm that the choice of forcing dataset, rather than the model architecture, remains the primary factor controlling predictive skill, with the high-resolution NLDAS-2 forcing yielding performance closest to the ParFlow-CONUS1 reference.

The ConvLSTM surrogate model was also assessed against independent in-situ observations from USGS groundwater monitoring wells (U.S. Geological Survey, 2024) to provide a more rigorous test of model transferability. Figure 3.8 presents the spatial distribution of Pearson (Spearman) correlation and RMSE between ConvLSTM-predicted and USGS-observed water table depth anomalies at the available well locations, together with representative point-scale time series comparisons.

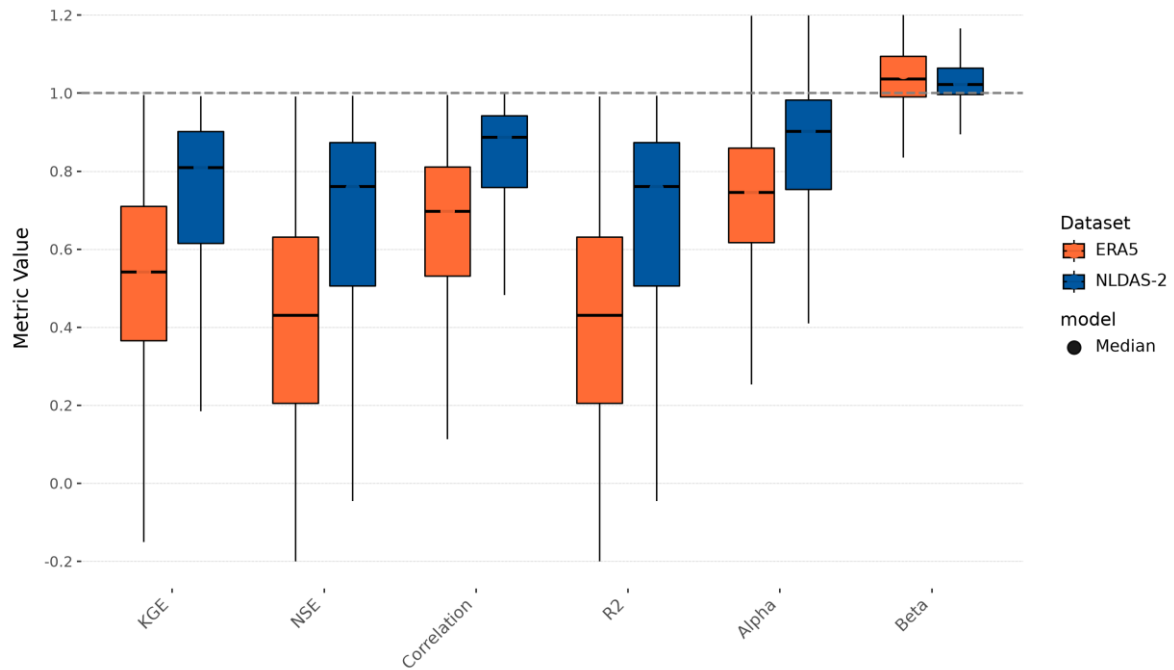


Figure 3.7. Summary performance of the ConvLSTM model in predicting daily Snow Water Equivalent (SWE), forced with NLDAS-2 and ERA5 reanalysis data, evaluated against ParFlow-CONUS1. Boxplots show the distribution of KGE, NSE, Pearson correlation, coefficient of determination (R^2), variability ratio (α), and bias ratio (β) across valid grid cells. The dashed line at 1.0 marks ideal model performance where applicable.

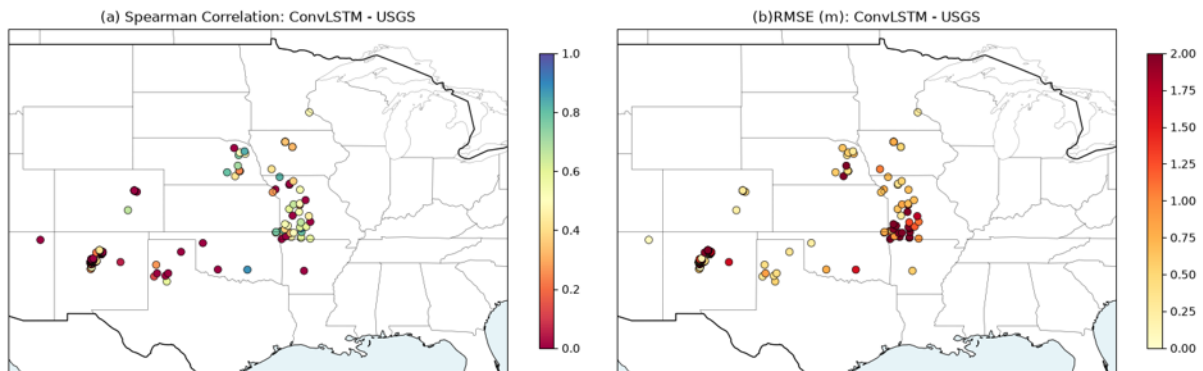


Figure 3.8. Evaluation of the ConvLSTM model for daily water table depth (WTD) anomalies against USGS in-situ observations over the U.S. study domain for the period 2003–2005.

The spatial maps of correlation and RMSE reveal a heterogeneous performance without a simple directional gradient. Pearson (Spearman) correlation values are predominantly positive across most monitoring locations, indicating that the ConvLSTM model successfully captures the temporal phasing of observed groundwater anomalies, broadly consistent with the moderate-to-good temporal skill reported by O'Neill et al. (2021). The highest correlations are concentrated in the dense central cluster of wells, where several sites exceed $p = 0.6$ – 0.8 . The lowest correlations, by contrast, are concentrated in the southwestern cluster of wells. The convergence of both models' weaknesses in this region suggests a shared structural limitation, rather than a surrogate-specific error, in reproducing groundwater dynamics where water tables are deep and heavily managed.

The RMSE map shows a partly distinct spatial pattern: the highest errors (up to ~2 m) occur precisely within the densely sampled central region, while the lowest errors are found at the isolated northeastern well. The point-scale time series confirm that the ConvLSTM model reproduces the sign and approximate timing of individual anomaly events at USGS monitoring locations, although the amplitude of predicted anomalies tends to be attenuated relative to observations.

It should be noted, however, that this evaluation is constrained to the 2003–2005 period, the only interval for which USGS well records overlap with the ParFlow-CONUS1 simulation window available through the HydroData portal; this three-year comparison window is short relative to the multi-year, low-frequency variability typical of groundwater systems, and it limits the extent to which the reported correlation and RMSE statistics can be considered representative of longer-term surrogate performance or of rarer, low-frequency recharge and drought events.

3.4. Conclusions

This chapter evaluated the transferability capacity of the ConvLSTM surrogate model, previously developed and trained on European data (Avila et al., 2026a), to the contiguous United States for the prediction of water table depth (WTD), without any retraining process. The ConvLSTM framework was applied, using WTD simulations from the physically based ParFlow-CONUS1 model as the reference dataset. Two forcing configurations were tested: the high-resolution NLDAS-2 reanalysis, which is used as forcing data in the ParFlow-CONUS1 model, and the coarser, globally available ERA5-Land reanalysis dataset. The ConvLSTM predictions were evaluated both spatially and temporally against the ParFlow-CONUS1 reference and further validated against independent in-situ groundwater observations from the USGS monitoring network for the 2003–2005 period, providing an additional, observation-based test of the surrogate's transferability beyond the simulation domain.

Results confirm that the framework retains meaningful predictive skill for water table depth in a hydroclimatically and topographically distinct domain. Forced with the native, high-resolution NLDAS-2 reanalysis, the model reproduces both the spatial structure and temporal dynamics of the ParFlow-CONUS1 reference. In general, model performance is better in the shallow, recharge-dominated water tables of the eastern Great Plains and remains reasonable in more challenging areas further west. Comparison against independent USGS in-situ observations confirms this result: the surrogate captures the sign and approximate timing of groundwater anomalies at monitoring wells across the domain, with skill broadly comparable to that of the ParFlow-CONUS1 reference itself against the same observational network; the corresponding evaluation of the numerical model against in-situ data is presented in O'Neill et al. (2021), to which the reader is referred for further details.

Some limitations can be discussed in this study. First, predictive skill degrades toward the central and western portions of the domain. Deep water-table conditions over major aquifers and the increasing topographic complexity pose a substantially greater challenge. Part of this degradation reflects known biases and reduced reliability of the ParFlow-CONUS1 reference itself in these regions, not only surrogate error alone. Second, forcing-data quality remains the dominant control on transferred performance. Replacing the native NLDAS-2 reanalysis with a globally available ERA5-Land product produces a clear, spatially coherent loss of skill, particularly in the shallow, precipitation-responsive aquifers, where high temporal-resolution forcing has more significant impact. Third, both the ConvLSTM surrogate and the ParFlow-CONUS1 reference systematically underestimate the amplitude of USGS-observed WTD anomalies. This attenuation is consistent with sub-grid averaging of aquifer heterogeneity and local recharge conditions inherent to the selected grid, and it is compounded by the shallow-bias tendency already documented for ParFlow-CONUS1 against filtered USGS wells.

A more fundamental limitation concerns the USGS validation itself: it is constrained to a short three-year overlap period (2003–2005) and a limited, filtered subset of wells. Groundwater systems typically exhibit multi-year, low-frequency variability, so a three-year window cannot capture longer-term recharge cycles, drought recovery, or rare extreme events. The correlation and RMSE statistics reported here should therefore be treated as indicative of short-term skill rather than as a complete measure of the surrogate's long-term reliability, and they are inherently less robust than the pixel-wise comparison against the ParFlow-CONUS1 reference, which covers the full simulation period.

4. Predictive mapping of groundwater-related variables

4.1. Introduction

The predictive mapping tool MLMapper 2.0 was developed by UCM as part of the STARS4WATER (Baron & Keller, 2026). The mapping tool offers a strong capacity for extrapolation to other river basins and study areas (Avila et al., 2026a). It is highly versatile because the architecture is not dependent on any specific database. MLMapper generates spatial predictions for any variable—here, groundwater-related variables—provided that a set of observations or ground-truth points is available from which a binary target variable can be defined. This variable may relate to groundwater quality, the presence of the resource, or the existence of groundwater-dependent ecosystems.

This capability has been explored and validated both within the context of the river basins included in the project (Duero and East Anglia) and in basins or regions from other parts of the world. The following sub-sections present the results, the specifications of each area, as well as the challenges encountered throughout the process of applying the method to other regions outside Europe.

4.2. MLMapper application in Chile

In this context, the proposed methodology was applied to the Ligua-Petorca basins in central Chile (Figure 4.1). The study provides a practical example of how this type of approach can be used to identify groundwater-dependent ecosystems. Identifying and mapping these ecosystems is essential for their effective management and conservation, a need underscored by recent global assessments showing that groundwater-dependent ecosystems are widespread yet largely unprotected in dryland regions (Rohde et al., 2024). The process combines field observations regarding the presence of these environments with a set of environmental predictor variables derived from remote sensing and geospatial data that influence the presence of these wetlands or ecosystems.

As mentioned earlier, MLMappers's versatility allows both the target and explanatory variables to be adapted to the specific needs involved in each case. The target variable was defined as the presence or absence of groundwater-dependent ecosystems (GDEs), based on a ground-truth dataset compiled from national wetland inventories and previous studies (Durán-Llaser et al., 2022). This dataset included a wide range of ecosystem types, such as grasslands, shrubs, springs, forests, riparian vegetation, and flooded wetlands, ensuring broad representativeness across the study area. A total of 3,067 training points were selected, balancing spatial coverage and data availability. The modelling framework incorporated a set of 21 explanatory variables, including geological, climatic, topographical, and satellite-derived factors, chosen based on an extensive literature review and their relevance for predicting groundwater-dependent ecosystems.

From a methodological perspective, the impact of sample size on model predictive performance was analyzed. Specifically, machine learning performance metrics were evaluated using 100%, 75%, 50%, and 25% of the dataset. The objective of this analysis was to assess the extent to which model performance depends on sample size to achieve reliable results (Figure 4.2).

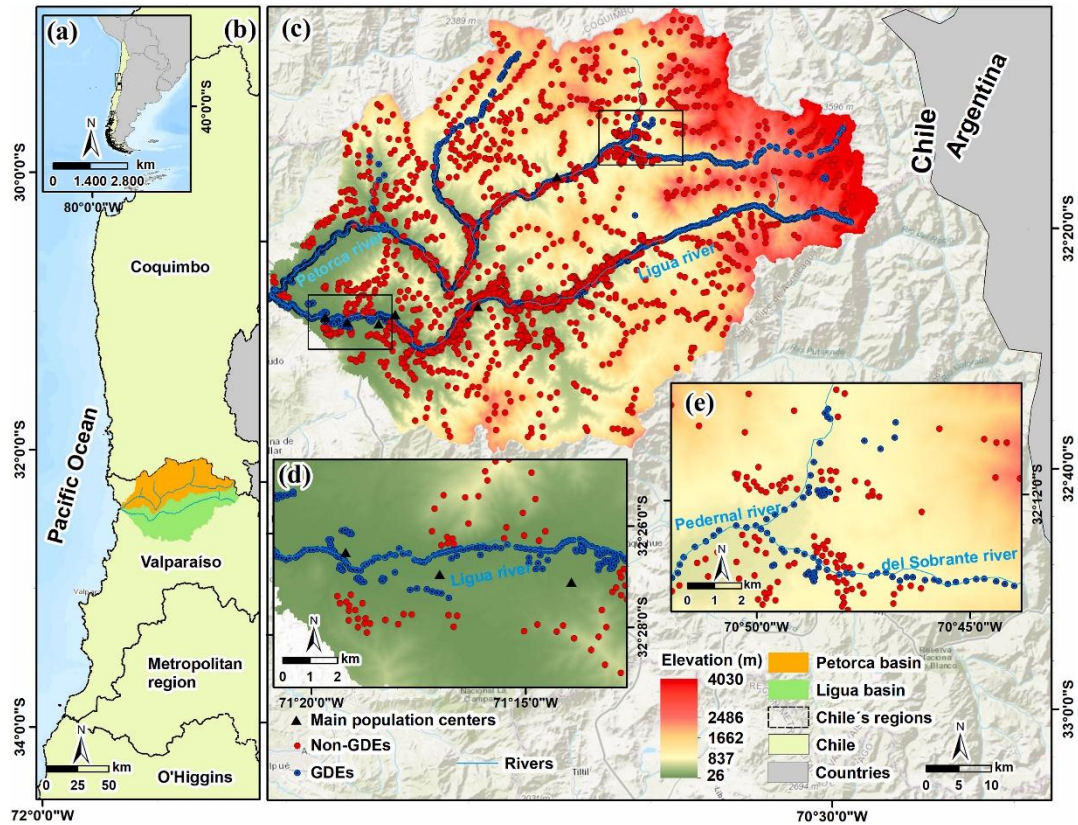


Figure 4.1. Geographic location and details of the Ligua and Petorca basins. a) geographic context within South America; b) Location within Chile and its regional divisions; c) Elevation and training points for groundwater-dependent ecosystems (GDEs), indicating presence and absence, and d-e) enlarged views of selected sectors highlighting the distribution of GDE training points. Obtained from Durán-Llaser et al. (2025).

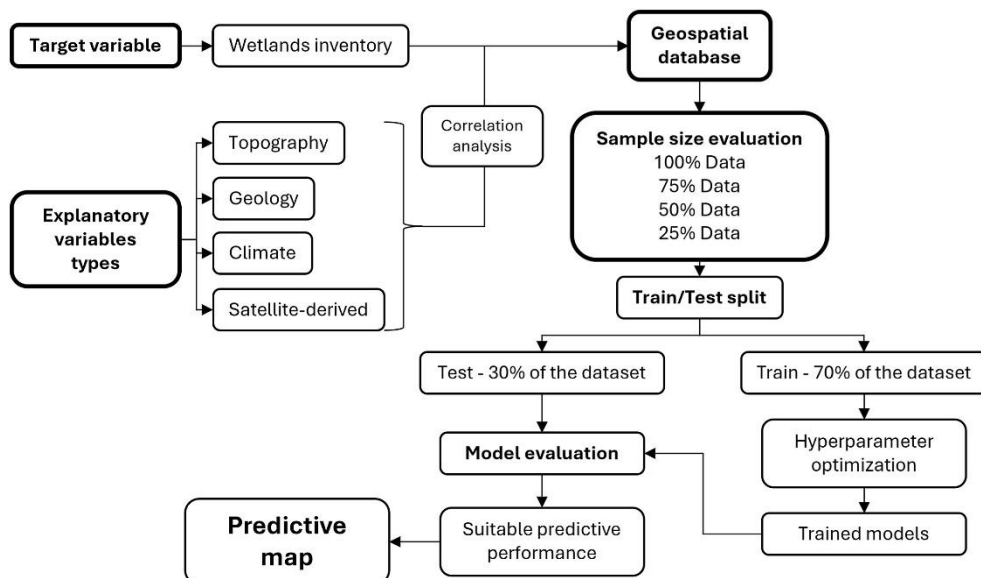


Figure 4.2. Schematic representation of the methodology used in the study, showing the sequence from the compilation of target and explanatory variable databases, through statistical analysis and model evaluation, to the generation of the final predictive map. Obtained from Durán-Llaser et al. (2025).

The results obtained highlight the effectiveness of this approach in capturing the spatial patterns associated with groundwater dependence. All models achieved scores above 0.90 across the test accuracy, F1 score for both classes, and the area under the curve (AUC), indicating a strong ability to predict the target variable (Table 4.1) and demonstrating the capacity of environmental variables to explain the occurrence of these ecosystems. Moreover, the comparison across different data subsets revealed that predictive performance remained largely stable despite reductions in sample size. Even when the dataset was reduced to 25% of its original size, only a negligible decline in performance metrics was observed.

Table 4.1. Summary of the machine learning metrics obtained in the different databases for the best performing models. All dataset is 3067 training points of presence or absence of GDEs, 75 % corresponds to 2300 randomly selected training points, 50 % are 1534 and 25 %, 767 points. Modified from Durán-Llaser et al. (2025).

Subsets	Metrics	Algorithm			
		RFC	ABC	GBC	ETC
ALL DATASET	Test score	0.965	0.961	0.976	0.961
	F1 score 0	0.970	0.960	0.980	0.960
	F1 score 1	0.960	0.960	0.970	0.960
	AUC	0.995	0.991	0.998	0.996
75 % DATASET	Test score	0.955	0.951	0.959	0.957
	F1 score 0	0.960	0.960	0.970	0.960
	F1 score 1	0.950	0.940	0.950	0.950
	AUC	0.991	0.988	0.991	0.993
50 % DATASET	Test score	0.965	0.944	0.963	0.954
	F1 score 0	0.970	0.950	0.970	0.960
	F1 score 1	0.960	0.930	0.960	0.950
	AUC	0.995	0.991	0.993	0.993
25 % DATASET	Test score	0.948	0.944	0.965	0.918
	F1 score 0	0.960	0.950	0.970	0.930
	F1 score 1	0.94	0.9400	0.960	0.900
	AUC	0.991	0.99	0.991	0.984

The predictive maps obtained illustrate the spatial distribution of groundwater-dependent ecosystems (GDEs) across the study area (Figure 4.3). Higher probabilities of locating GDEs were consistently associated with river valleys and flat areas, where strong surface–groundwater interactions occur, regardless the sample size used. Notably, areas near the outlets of the Ligua and Petorca rivers, and the coastal wetlands, showed a significant presence of GDEs, characterized by permeable substrates, high humidity, dense vegetation, and gentle slopes. In contrast, low-probability zones corresponded to steep and structurally complex terrains, as well as agricultural areas. The consistent classification of valley-bottom agricultural land as very low GDE potential is particularly relevant, as it enables accurate delineation of GDEs while excluding human-dominated landscapes. Although all scenarios showed strong performance, models trained with 75% and 100% of the data aligned more closely with expert knowledge, with the full dataset providing the most representative results.

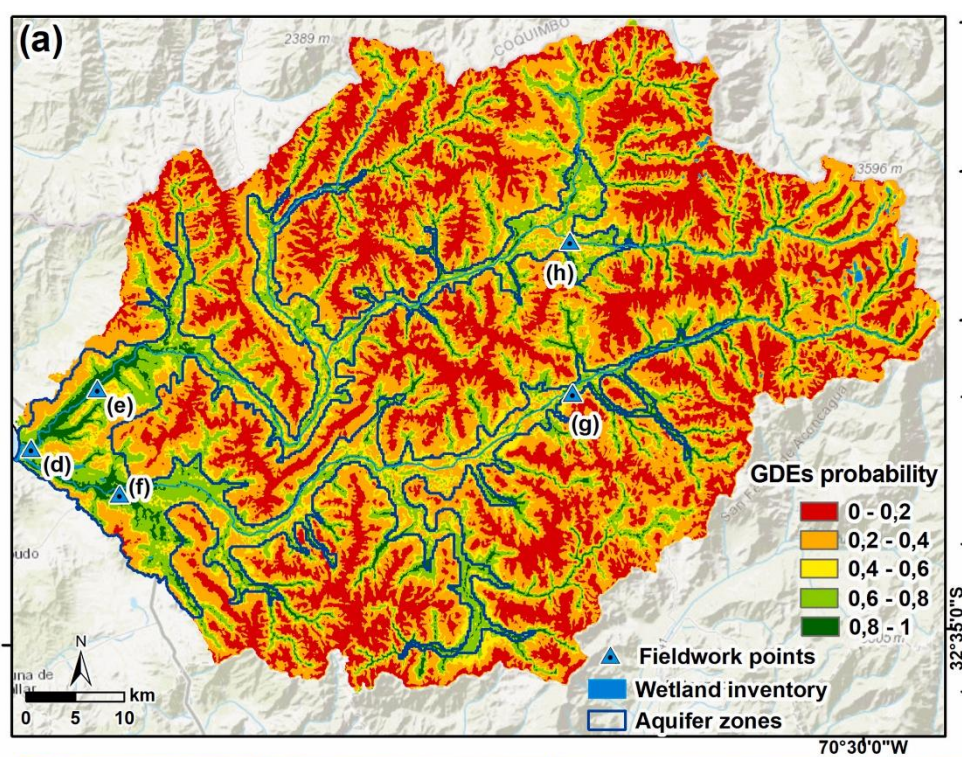


Figure 4.3. GDEs probability or potential in the Ligua and Petorca basins. Represents the GDEs probability or potential, ranging from 0 (low) to 1 (very high), wetland inventory by MMA, aquifer zones by DGA and some fieldwork points by Duran-Llacer et al. (2022). Modified from Durán-Llacer et al. (2025).

4.3. MLMapper application in Madagascar

The application of predictive mapping in Madagascar has a different objective compared to the one described above and the original application within the STARS4Water project. Access to safe drinking water remains a critical challenge in many regions of the world, particularly in sub-Saharan Africa, where large portions of the population still lack reliable and good-quality water sources. In southern Madagascar, this situation is exacerbated by arid conditions, recurrent droughts, and strong dependence on groundwater as the primary source of freshwater. However, groundwater availability and quality are highly variable and increasingly threatened specially by saline intrusion. In this context, predictive mapping approaches based on machine learning offer

a valuable tool to identify areas with suitable groundwater quantity and quality. The use of MLMapper in this context aims to identify areas where solar-powered pumps could be installed, which requires identifying high-yield areas along with quality standards, supporting the identification of sustainable water resources and improving decision-making in vulnerable regions.

The first step in this project involved analysing pumping tests from more than 1,200 locations across southern Madagascar. This analysis, based on step-drawdown pumping tests, made it possible to identify, using various methodologies, the locations that offered optimal conditions, in terms of well yield and hydrogeological conditions, for the installation of a solar pump requiring a minimum flow rate of 2 m³/h, although a rate of 4 m³/h is preferable (Fussi et al., 2025). In this regard, boreholes with a minimum test flow rate of 1.65 m³/h in which the efficiency and the drop to 2 m³/h were 60% and less than 5 meters, respectively, were considered positive. In addition, once the optimal points were selected based on yield, they were evaluated against quality criteria. For this purpose, electrical conductivity values were used, given that the high salinity of local groundwater represents a major regional issue that prevents its use for human consumption. The threshold of 2000 µS/cm was considered. The spatial distribution of the positive and negative points in terms of yield and quality is shown in Figure 4.4 and figure 4.5.

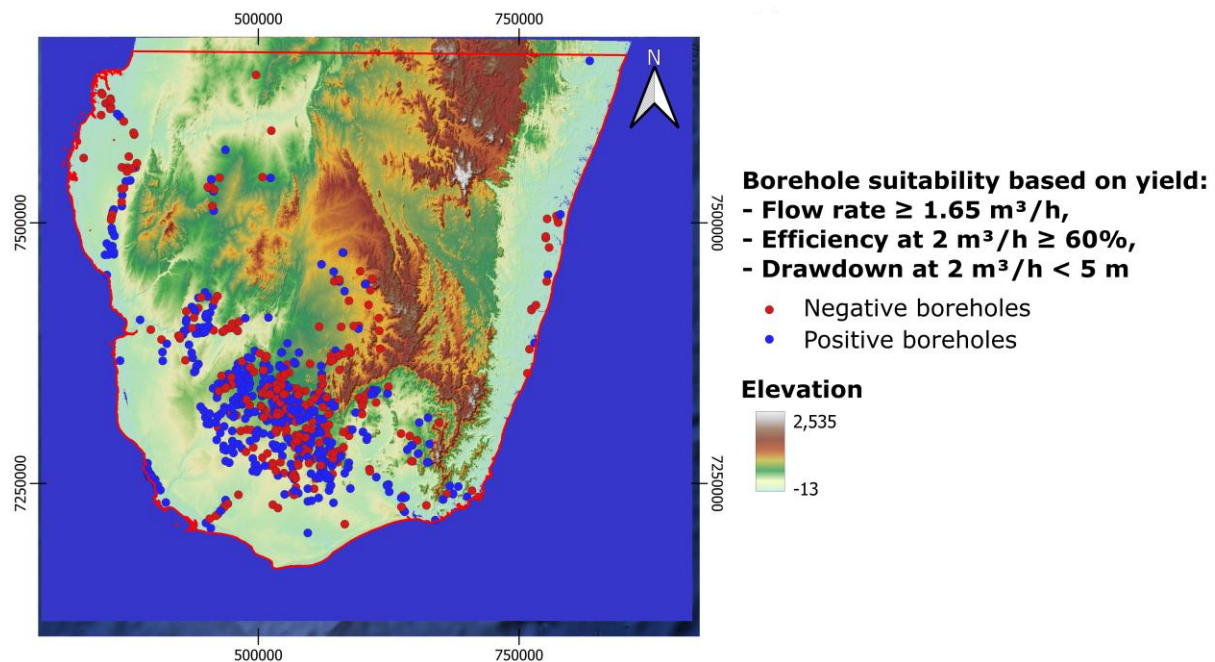


Figure 4.4. Spatial distribution of the positive and negative boreholes in southern Madagascar based on yield.

The predictive performance of the models, as measured by standard machine learning metrics, was lower than that of previous studies. In the case of predictive quality mapping, the metrics were around 0.84 for accuracy, 0.78 for AUC, and 0.82 for F1 score. Meanwhile, in the approach based on borehole performance, the values obtained were around 0.70 for all three metrics analysed.

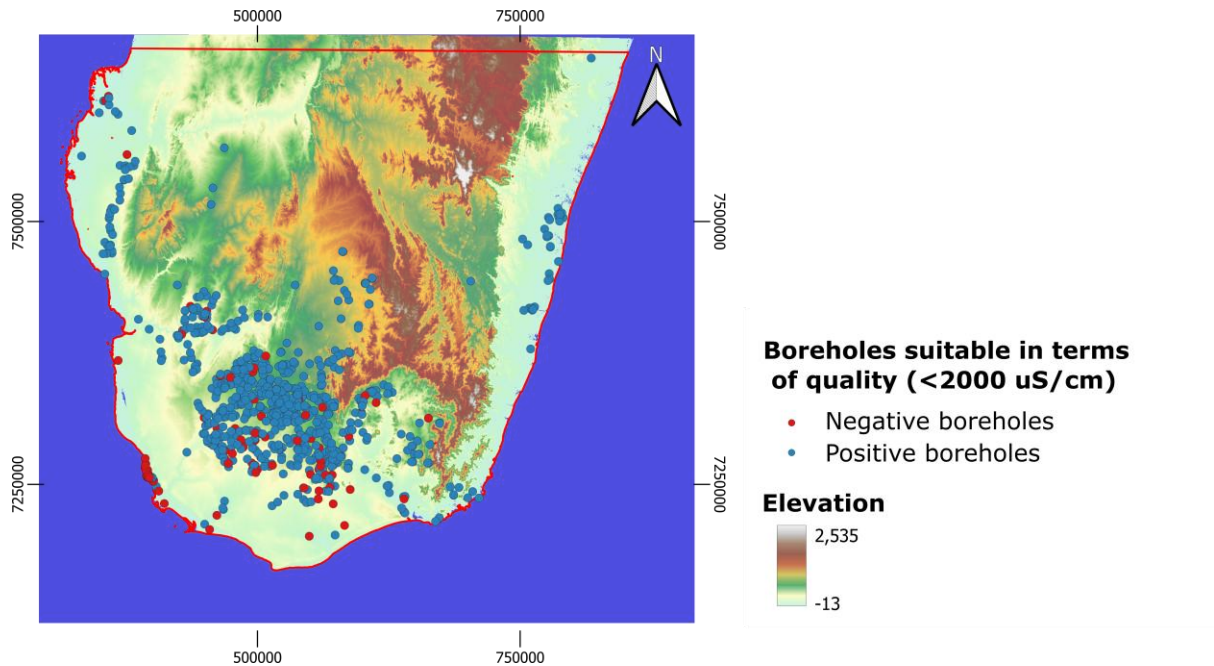


Figure 4.5. Spatial distribution of the positive and negative boreholes in southern Madagascar in terms of quality.

Figure 4.6 and Figure 4.7 show spatial predictions for both approaches. The blue areas on both maps correspond to suitable areas for groundwater exploitation, either in terms of quality or quantity. This means that the machine learning algorithms have identified a combination of explanatory variables that is most likely to yield favourable conditions based on the ground truth data used for training.

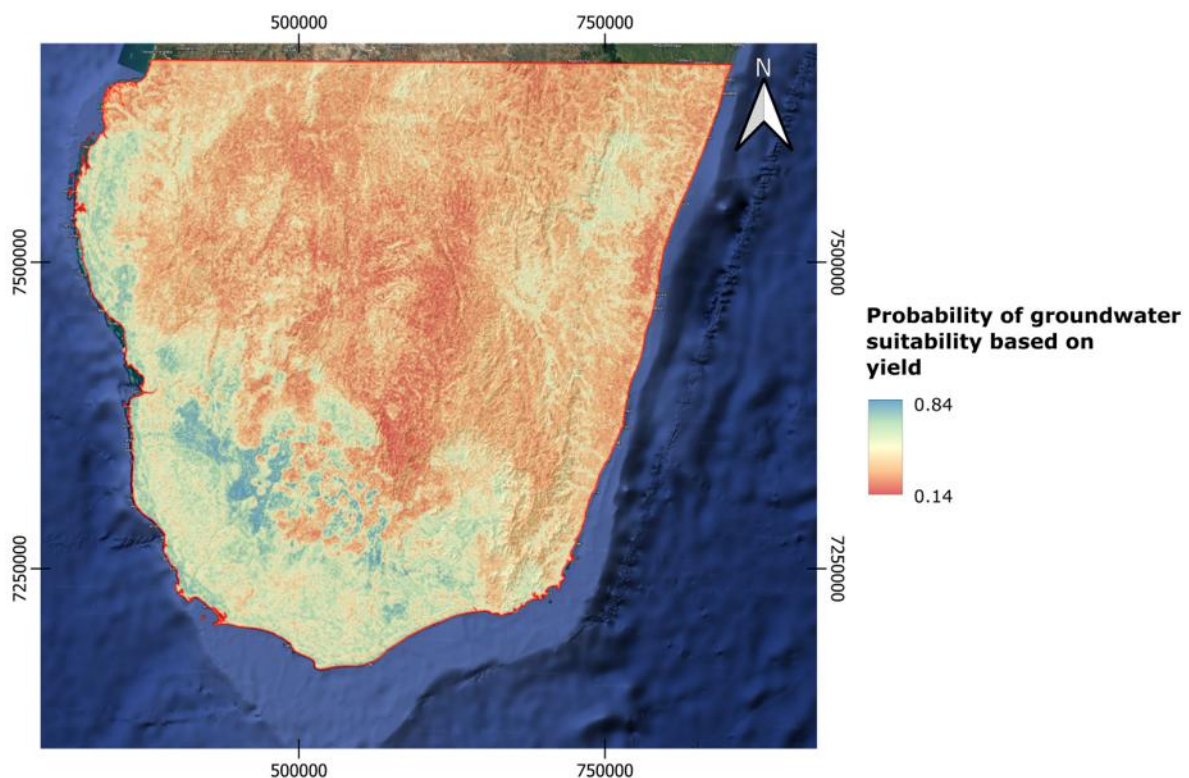


Figure 4.6. Spatial prediction of groundwater exploitation suitability based on yield.

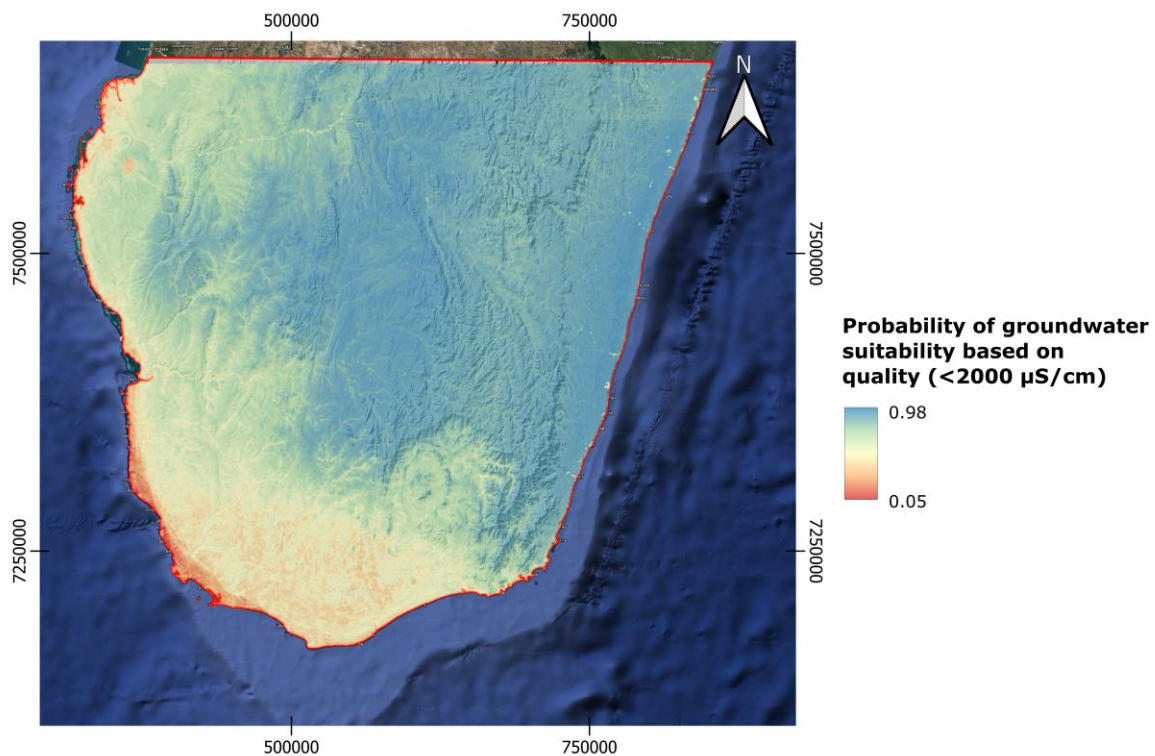


Figure 4.7. Spatial prediction of groundwater exploitation suitability based on quality.

4.4. Conclusions

The borehole database presents several limitations that affect the quality and representativeness of the analysis. Firstly, the limited number of reliable pumping tests, due to low yields or outdated measurements, has led to the partial exclusion of relevant data, thereby reducing the effective size of the dataset. In addition, the explanatory variables used in the study are constrained by limited spatial resolution, which hinders the ability to capture important local variations in hydrogeological conditions. This issue is further compounded by the lack of subsurface information, such as faults and structural lineaments, whose absence reduces the overall reliability of the predictions.

Moreover, the application of machine learning methods is also constrained by data availability and quality. The relatively small dataset limits the robustness of the model and its capacity to generalize to new areas. There is also significant spatial uncertainty in regions that are poorly represented in the dataset, meaning that predictions in these areas should be interpreted with caution. Altogether, these limitations highlight the need to complement the results with field data and detailed hydrogeological investigations.

Groundwater is a crucial resource in arid regions of developing countries, where a large proportion of the population relies on boreholes and excavated wells for drinking water supply. In this context, groundwater exploitation suitability maps developed through the integration of GIS and machine learning techniques represent a valuable tool to support large-scale exploration campaigns, helping to identify areas with higher groundwater potential. However, this approach should not be considered a low-cost substitute for traditional exploration methods, but rather as a complementary tool to prioritize the most promising areas for targeted hydrogeological surveys. In addition, the implementation of solar-powered pumps in existing boreholes could significantly improve water access in regions such as southern Madagascar, offering a sustainable and context-appropriate solution in areas with limited energy infrastructure.

5. Ensemble-tree models for reservoir storage

5.1. Introduction

In STARS4Water, an ensemble-tree (ET) model for forecasting reservoir storage at a monthly timestep was developed using data from 15 reservoirs in the Duero basin and 14 over the UK, and extended to include 4 reservoirs from the Seine basin (D3.4) (Baron & Keller, 2026). This model was designed to help improve basin-wide assessments of water resources at seasonal to sub-seasonal timescales, an area identified by stakeholders. Full details on the model, and analysis of performance for the European reservoirs, can be found in Baron & Keller (2026).

Most data-driven models for reservoir forecasting are single-site models (i.e. trained and applied at one reservoir), although Rao et al. (2026) explore both deep learning and a Random Forest approach to forecast outflow at two reservoirs. This contrasts with streamflow prediction, where multi-basin models are increasingly common (Kratzert et al., 2019; Lees et al., 2021). Multi-site models have the potential for improved model performance, particularly for extreme events, due to the greater number of extreme events seen in the training period compared to a single-site model (Kratzert et al., 2024; Fang et al., 2022). However, this trend is not universal: Tran et al (2025) show that single-site streamflow models have superior performance compared to a global model in 46% of case study catchments in the US and UK.

The model developed in Baron & Keller (2026) is a multi-reservoir model, combining timeseries data from 33 reservoirs over the UK, Spain and France. This approach has been taken to improve model performance, since the data available at each site is limited (particularly due to the monthly timestep), and allows the inclusion of reservoirs with shorter timeseries. This work extends the model to include reservoirs in Thailand and is supported by the National Capacity for Global Challenges Programme¹, alongside the STARS4Water project.

A range of model set-ups were explored, comparing single-reservoir models to a regional model (trained only on Thai reservoirs) and a 'global' model (trained in the Thai and European reservoirs). Leave-one-out analysis was also conducted to explore how multi-site models contribute to model performance for each Thai reservoir. This analysis provides information for developing a robust training dataset for any modeller planning to apply a similar method in their region.

5.2. Region and Data

Thailand is a tropical monsoon country in southeast Asia, approximately 523,000 km². The population has historically been primarily rural but is rapidly urbanising; the rural population was 81% of total population in 1990, compared to 38% in 2024 (World Urbanization Prospects, accessed 2026). However, agriculture is still a dominant water user, accounting for over 90% of total water abstracted in 2023 (FAO, 2026). There are a number of large reservoirs within Thailand (see Figure 5.1 and Table 5.1), with total capacity estimated at 32% of annual runoff. Many of these reservoirs are multi-purpose, with irrigation, water supply and hydropower being dominant uses.

¹ NERC – NE/X006247/1.

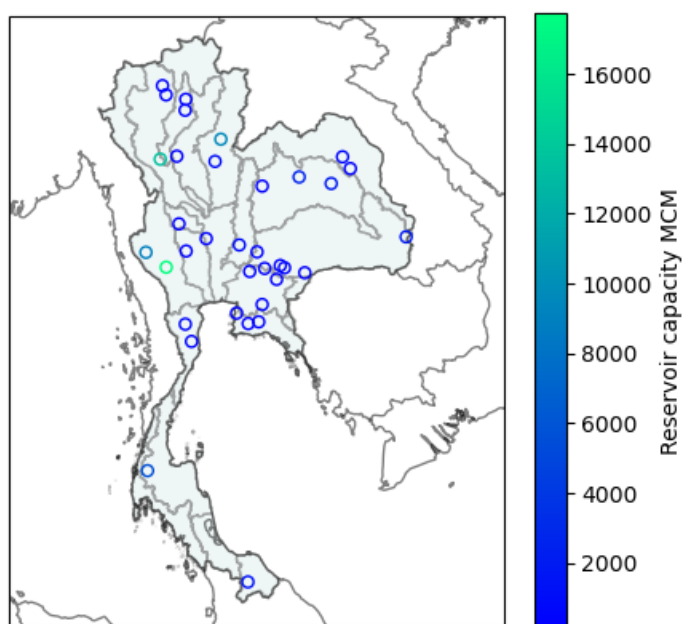


Figure 5.1. Thailand, with major basins outlined in grey and major reservoirs (circle markers).

Table 5.1. Thai reservoirs included in the study.

Name	Code	Latitude	Longitude	Capacity (MCM)	Catchment Area (km ²)	Purpose *
Mae Kuang Udom Thara	MKU	18.93	99.12	263	558	Irr
Kiew Lom	KLM	18.52	99.62	106	2747	Irr; WS
Kiew Kor Mah	KKM	18.81	99.64	170	1275	Irr; P
Khwae Noi	KNI	17.18	100.41	939	4254	Irr; WS; P; FC
Mae Mok	MMK	17.32	99.41	110	765	Irr; WS
Bhumibol	BMB	17.24	98.97	13462	26214	Irr; WS; P; FC
Sirikit	SRK	17.76	100.56	9510	13203	Irr; WS; P; FC
Mae Ngat Sombun Chon	MNS	19.17	99.03	265	1274	Irr; P; FC
Huai Luang	HLG	17.36	102.57	136	665	Irr; WS; FSH

D5.3: UPSCALING MACHINE LEARNING MODELS TO OTHER RIVER BASINS

Nam Un	NUN	17.30	103.76	520	1087	Irr
Lam Pao	LPO	16.60	103.46	1980	6061	Irr; WS
Lam Takhong	LTK	14.81	101.51	314	1331	Irr; WS
Lam Phra Phloeng	LPP	14.59	101.84	155	842	Irr; WS
Munbon	MBN	14.46	102.12	141	552	Irr; WS
Lam Chae	LCE	14.39	102.23	275	683	Irr
Lam Nang Rong	LNR	14.26	102.77	121	457	Irr; WS
Nam Pung	NPG	16.99	103.96	165	286	Irr; P; FC
Chulabhorn	CLB	16.54	101.65	164	516	Irr; P
Ubolratana	UBR	16.77	102.62	2431	12117	Irr; WS; P; FC; FSH
Sirindhorn	SRD	15.21	105.43	1966	2051	Irr; P
Pasak Chonlasit	PSC	14.99	101.05	960	12970	Irr; P; FC
Tabsalao	TSL	15.55	99.46	160	531	Irr; WS
Krasiao	KSO	14.83	99.65	299	1230	Irr; FC
Srinakarind	SNK	14.41	99.13	17745	10999	Irr; WS; P; FC
Vajiralongkorn	VJR	14.80	98.60	8860	3695	Irr; WS; P; FC
Khun Dan Prakan Chon	KDP	14.30	101.32	224	202	Irr; P
Khlong Si Yat	KSY	13.43	101.65	420	937	Irr
Bang Pra	BPA	13.21	100.98	117	79	Irr; WS
	NPL	12.93	101.28	164	422	Irr; WS
Pra Sae	PSE	12.98	101.56	295	598	WS
Naruebod- indrachinta	NRB	14.09	102.03	295	384	Irr; FC
Pran buri	PBR	12.46	99.79	391	1991	Irr; FC

Kaeng Krachan	KKC	12.92	99.63	710	2181	Irr; P; FSH
Ratchaprapha	RCP	9.08	98.64	5639	1371	Irr; WS; P; FC
Bang Lang	BLG	6.16	101.27	1454	1978	Irr; P; FC

***Irr: Irrigation; WS: Water Supply; P: Hydropower; FC: Flood Control; FSH: Fisheries**

This work extends the dataset collated in D3.4 (Baron & Keller, 2026). Timeseries data for reservoir storage were retrieved from the Royal Irrigation Department (RID, 2024). Meteorological data was derived from the ERA5 dataset (Muñoz Sabater, 2019) and catchment characteristics from HydroATLAS (Linke et al., 2019), following the methods of Kratzert et al. (2023). Reservoir characteristics were derived from the Grand database (Lehner et al., 2011).

5.3. Method

Three types of ET model were produced:

- a global model, including data from all reservoirs (Thai and European)
- a regional model, including just Thai reservoirs
- single-site models for each reservoir with sufficient data

For the single-site models, timeseries of historic reservoir storage and catchment rainfall and temperature were used to train the model. For the regional and global models, static catchment and reservoir characteristics were also used. Each model build followed the steps outlined in D3.4 (Baron & Keller, 2026).

A leave-one-out analysis was conducted on the regional model to investigate how model performance for each Thai reservoir is influenced by the reservoirs present in the training set (henceforth referred to as donor reservoirs). This was undertaken by producing a set of model performance contributions for all combinations of target and donor reservoir pairs, calculated as:

$$\varphi_{d \rightarrow t} = NSE_{r-t} - NSE_{r-t-d}$$

Where $\varphi_{d \rightarrow t}$ is the contribution for target reservoir t from donor d , and $NSE_{r-t, r-t-d}$ is the Nash-Sutcliffe Efficiency metric for the target reservoir test period for a model trained on the regional dataset minus the target reservoir ($r-t$) and the regional dataset minus the target and donor reservoirs ($r-t-d$).

A meta-model is then built to see whether $\varphi_{d \rightarrow t}$ can be predicted using reservoir or catchment characteristics. These characteristics include hydro-, geo- and meteorological aspects, as well as geographic proximity, reservoir purpose and capacity, and various time-series metrics (high/low storages, seasonal vs residual components).

Finally, a dataset building experiment was carried out to determine how the number of reservoirs in the training set impacts model performance. For this, the donor pool was initiated with the target reservoirs and new reservoirs were added at random (with 8 shuffles), with model training and evaluation after each addition.

5.4. Results

Each model was evaluated for the Thai reservoirs for the test period of each single model, using the Nash-Sutcliffe Efficiency metric (NSE), Root Mean Square Error (RMSE), and PBIAS, see Table 5.2. Timeseries of observed and simulated storages are depicted in Figure 5.2 for a set of reservoirs demonstrating different levels of performance.

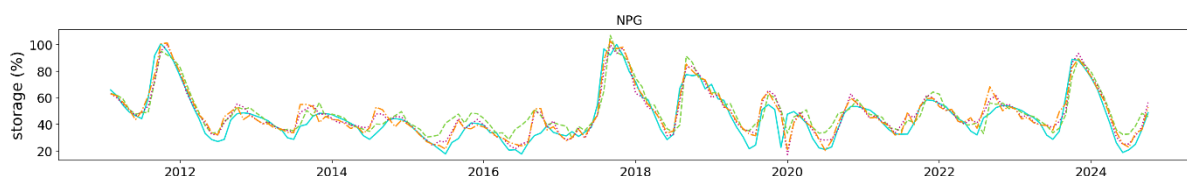
The results demonstrate that, while a regional model generally outperforms a single-reservoir model (higher NSE on 26 of 35 reservoirs), a global model does not add value compared to a regional model (see Table 5.2). This improvement over the single-reservoir models is often due to a better simulation of more extreme low and high storages, for example the low-storage periods in 2015, 2016 and 2024 in NPG and the high-storage periods in 2018, 2022 and 2023 in KSO (Figure 5.2).

Table 5.2. Model performance for the Thai reservoirs over the testing period, estimated using the Nash-Sutcliffe Efficiency (NSE), Root Mean Square Error (RMSE) and PBIAS, for single-site (S), regional (R) and global (G) models.

Reservoir code	NSE			RMSE (%)			pbias (%)		
	S	R	G	S	R	G	S	R	G
BLG	0.75	0.79	0.73	8.0	7.3	8.3	0.3	-0.4	-5.4
BMB	0.93	0.94	0.91	4.2	4.0	4.8	-1.1	0.1	-1.7
BPA	0.86	0.89	0.89	8.1	7.1	7.3	0.8	0.2	0.2
CLB	0.89	0.92	0.92	7.5	6.4	6.1	-2.4	-3.0	-2.4
HLG	0.90	0.92	0.92	6.7	6.0	5.9	2.3	-0.9	-1.1
KDP	0.90	0.93	0.94	9.2	8.1	7.3	7.4	-0.4	0.5
KKC	0.70	0.84	0.79	8.7	6.4	7.4	-0.1	-4.1	-3.0
KKM	0.90	0.92	0.91	8.4	7.5	8.0	2.3	2.7	2.2
KLM	0.68	0.76	0.75	10.5	9.1	9.3	6.4	0.8	-0.3
KNI	0.81	0.92	0.92	12.0	7.7	7.9	9.4	2.7	1.5
KSO	0.88	0.94	0.93	8.9	6.4	6.8	4.4	0.6	-0.2
KSY	0.78	0.91	0.87	9.3	5.8	7.0	-6.2	-9.8	-13.4
LCE	0.93	0.92	0.92	6.2	6.3	6.4	2.0	0.1	-0.3
LNR	0.77	0.97	0.95	12.4	4.8	6.1	4.0	1.1	0.6
LPO	0.83	0.91	0.91	8.7	6.3	6.4	7.9	-0.4	-0.5
LPP	0.75	0.96	0.94	13.5	5.5	6.7	13.7	1.2	2.4

D5.3: UPSCALING MACHINE LEARNING MODELS TO OTHER RIVER BASINS

LTK	0.91	0.94	0.93	7.3	6.0	6.3	-0.2	1.5	-0.2
MBN	0.84	0.87	0.84	9.0	8.1	8.8	1.6	1.8	2.2
MKU	0.91	0.93	0.89	6.4	5.4	6.9	5.3	-0.1	-3.6
MMK	0.52	0.75	0.79	17.8	12.8	11.7	15.6	5.7	6.4
MNS	0.93	0.95	0.95	7.3	5.9	6.2	0.4	-0.1	-0.8
NPG	0.77	0.86	0.86	8.7	6.9	6.9	-8.0	-4.9	-5.0
NPL	0.81	0.85	0.84	10.7	9.6	9.8	3.6	3.1	2.9
NRB	0.88	0.92	0.94	8.9	7.0	6.4	5.1	2.7	1.8
NUN	0.93	0.90	0.90	5.3	6.3	6.2	0.1	-0.8	-0.3
PBR	0.81	0.88	0.87	8.2	6.5	6.7	-4.5	-2.5	-5.0
PSC	0.77	0.84	0.86	14.7	12.2	11.3	-13.1	-7.9	-2.4
PSE	0.86	0.95	0.94	8.4	5.2	5.5	3.6	1.0	1.8
RCP	0.64	0.59	0.30	5.8	6.2	8.1	-0.4	-0.6	-4.6
SNK	0.65	0.36	0.34	5.5	7.5	7.6	-0.3	0.9	0.6
SRD	0.83	0.75	0.77	6.3	7.6	7.4	-2.0	-1.5	-1.0
SRK	0.88	0.87	0.85	4.7	4.9	5.3	0.0	-0.8	-1.6
TSL	0.82	0.87	0.84	10.0	8.5	9.4	5.6	-1.9	-0.3
UBR	0.88	0.91	0.90	8.6	7.5	7.8	0.3	1.1	0.4
VJR	0.82	0.75	0.70	6.3	7.4	8.2	-0.3	-0.1	-3.1
Median	0.83	0.91	0.89	8.4	6.5	7.0	0.8	-0.1	-0.3



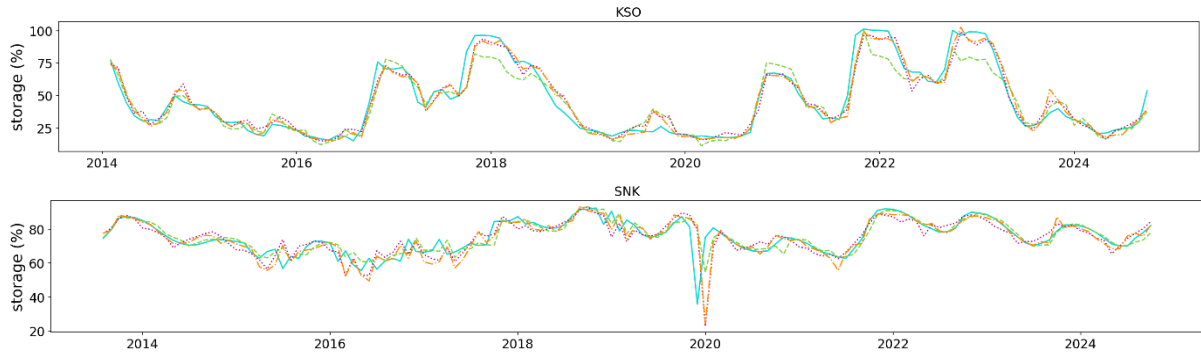


Figure 5.2. Observed and simulated monthly storage for case-study reservoirs. Solid blue: observed, dashed green: single model (S), dotted pink: regional model (R), dot-dash orange: global model (G).

Reduced performance can be caused by timing issues in extreme storages, for example, the low simulated storage for SNK in early 2020 where the regional and global models are a better match for the minimal value compared to the single model but have a month delay compared to the observed storage.

Figure 5.3 compares the relative NSE of the regional and single models for all the Thai reservoirs, and demonstrates that larger reservoirs, and reservoirs with longer time-series, tend to benefit less from a multi-site model (Spearman correlation -0.41 and -0.44 respectively). Reservoirs with a greater capacity are rarer, and often more highly managed than smaller reservoirs, and thus generally benefit less from a regional model.

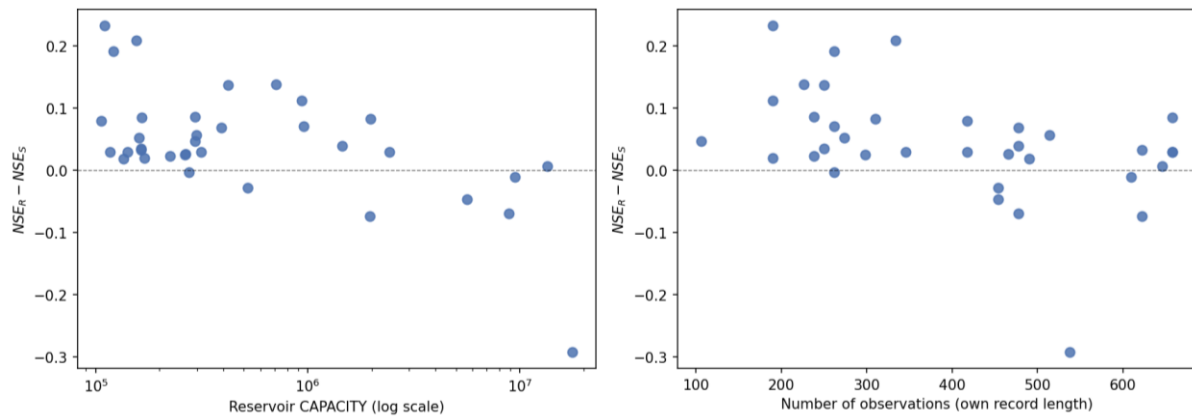


Figure 5.3. Relative model performance for the regional (R) and single (S) models compared to reservoir capacity and number of observations.

Leave-one-out analysis finds that the usefulness of the donor reservoirs to each target reservoir is generally independent of their characteristics, likely due to the general homogeneity of reservoir storage timeseries in this region. Even including the target reservoir in the dataset yields a median gain of only 0.015 NSE, although this effect is not evenly distributed: RCP and SNK both benefit greatly from being included (these two reservoirs see the most significant performance loss for the global model compared to their single-site models).

Figure 5.4 demonstrates that 90% of the model improvement associated with a larger number of reservoirs in the dataset can be achieved with around 9 additional reservoirs in the dataset.

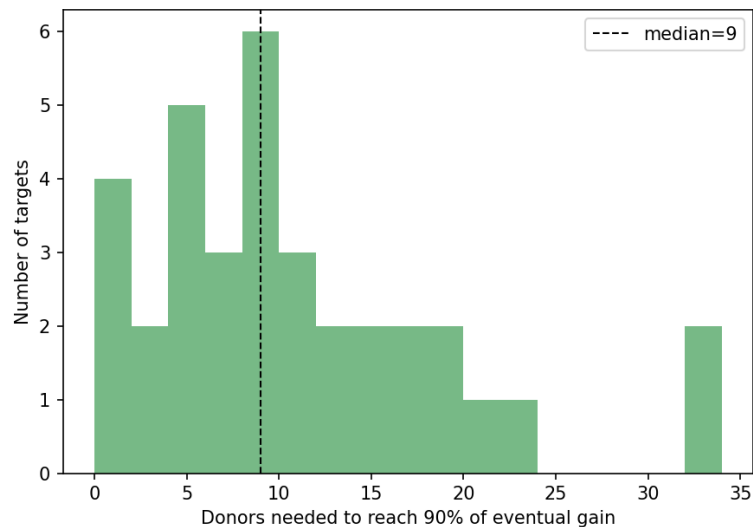


Figure 5.4. Histogram for the number of donor reservoirs required to achieve 90% of maximum model improvement.

5.5. Conclusions

This analysis compares ensemble-tree model performance for three model setups simulating reservoir storage at a monthly timestep: single-reservoir models (training only on the local reservoir), a regional model (training on reservoirs from Thailand), and a ‘global’ model (training on reservoirs from Thailand, Spain, France and the UK).

The results demonstrate that while a regional model generally outperforms a single-site model, due to the increase in information seen at training time, a global model often does not add value. This demonstrates that, while more data is often useful, the relevance of the data determines how important it is to data-driven methods, and there is a trade-off between valuable data and additional noise caused by a larger dataset. This indicates that a regional dataset would be sufficient for application of this method to other areas, although the value of additional data in a global model seems more beneficial to reservoirs with shorter time-series. There was no discernable reservoir characteristic that made its inclusion in the dataset particularly useful for model performance in Thailand, although in future work this could be explored further by including reservoirs from a greater range of countries. It is also worth noting that reservoirs subject to strong operational management, where releases are driven primarily by human decisions rather than hydrometeorological conditions, are inherently more difficult to represent using climate-based predictors alone, and this factor may partly explain the variability in performance observed across sites.

6. General conclusions and recommendations for wider uptake

The preceding chapters evaluated how well specific data-driven models perform when they are transferred from one region, climate, or data context to another, supporting a wider uptake of the developed models beyond Europe. The four case studies illustrate how different machine learning approaches can be applied with and without a full retrained procedure to extend the predictive ability, supporting water management.

For instance, a surrogate ConvLSTM model for daily Snow Water Equivalent can provide seasonal perspectives on snow storage that are relevant to hydropower operators, drought preparedness and flood awareness in mountainous basins. A surrogate ConvLSTM model for Water Table Depth offers groundwater managers fast, repeated state estimates at spatial resolutions that would be computationally expensive with process-based models alone. Predictive mapping helps environmental authorities delineate groundwater-dependent ecosystems for protection and land-use planning, and, in a development context, helps identify where sufficient groundwater quantity and quality is likely to be found, guiding investments in boreholes and solar pumping. Finally, ensemble-tree models of reservoir storage give reservoir operators and water allocation authorities seasonal views of storage evolution that inform release and abstraction strategies.

In all cases, the value of the tool lies in decision support: indicating where to look, when to act, and which management options are plausible, rather than replacing the judgement of the practitioner.

Viewed from a stakeholder's perspective, the general advantages of data-driven tools are considerable:

- **Portability:** A model developed once, centrally and at significant effort, can be adapted to new basins at comparatively low cost, so the benefit of a single investment multiplies across regions.
- **Lower entry barrier:** Many data-driven tools can be driven with globally available meteorological datasets and existing large-scale simulations, making them accessible even in basins that lack dense monitoring networks or calibrated process-based models.
- **Speed and flexibility:** They allow rapid evaluation of alternative scenarios and questions, and their modular design enables the same methods to be redeployed to different variables when regional priorities change.
- **Complementarity:** Rather than competing with established process-based models, observations, and stakeholder knowledge, they complement these elements within integrated modelling frameworks.
- **Communication value:** Their outputs translate naturally into maps, indicators, and time series that can be embedded in dashboards and story maps co-designed with stakeholders, easing communication between technical teams and decision makers.

At the same time, these tools have general limitations that stakeholders should understand before relying on them. Data-driven models learn from the data they are given, so they inherit the biases and blind spots of the simulations or observations used for training and cannot be expected to outperform the quality of their ground truth. In addition, their skill depends strongly on the quality of the input data; coarse or inconsistent forcing degrades predictions regardless of the sophistication of the method. Their behavior is not always transparent, and for many practitioners they retain a black-box character, while formal uncertainty quantification remains uncommon. They require validation against independent local observations before they can be trusted operationally in a new setting, and they should not be applied blindly outside the range of conditions for which they were built.

Furthermore, unlike many conventional process-based tools, data-driven models may require periodic retraining or recalibration as new data become available or as conditions evolve beyond those represented in the original training period, a task that is not always straightforward for practitioners who use the tool but were not involved in its original development.

The decision regarding adopting this type of tools should start from a clearly defined management question, not from the tool itself. High-quality, high-resolution local input data should be secured early, since this usually yields more benefit than any modification of the model. Training data should be curated for relevance to the target region rather than volume, and incremental adaptation, from direct transfer to targeted fine-tuning, should be preferred over full redevelopment. Once transferred to a new basin, every application should be validated against local observations that were not used during model development, confirming that its performance holds before it is used operationally. Stakeholders should be involved from the outset through co-design, so that the products address their real questions and are presented through channels they use, such as dashboards and story maps.

In the broader context of STARS4Water, whose ambition has been to create a new generation of data services and data-driven models tailored to stakeholder needs, these conclusions suggest that the tools demonstrated in this report show clear potential to move from research prototypes towards components of routine practice in river basin management, although the extent of stakeholder co-creation and uptake achieved so far varies across cases and would benefit from further clarification. If the conditions outlined above are met, the experience accumulated across Europe, North America, and other regions can be extended to other basins, including those with limited modelling tradition. However, transferability to basins with sparse monitoring networks should be considered with caution, since data availability itself is often a key constraint for the successful application and transfer of machine learning models, as illustrated by the reservoir storage case. Where this constraint can be addressed, such extension can support adaptive, resilient, and sustainable water management and contribute to EU policy objectives.

7. Bibliography

- Avila, L. & Kollet, S. (2025). *Multi-scale model integration: Improving existing models based on multi-scale model integration*. Deliverable 3.3. Version 1.2. STARS4Water Project. Horizon Europe Grant Agreement No 101059372.
- Avila, L., Aguilera, H., Gomez, V., Martinez, P., Baron, H., Engeland, K., Gelati, E., Hisdal, H., Hegdahl, T., Kollet, S., Salamon, P., Grimaldi, S., & Bisselink, B. (2026a). *Future of Europe's freshwater resources. Upscaling of basin-wide assessment of climate change risks and impacts to EU-level*. Deliverable 4.4. STARS4Water project. Horizon Europe Grant Agreement No 101059372.
- Avila, L., Engeland, K., Hegdahl, T. J., & Kollet, S. (2026b). A deep-learning surrogate based model to predict daily snow water equivalent in Southern Norway. *Journal of Integrated Water Basin Management* (manuscript submitted for publication).
- Baron, H. & Keller, V. (2026). *Data-driven modelling tools for STARS4Water river basins*. Deliverable 3.4. Version 1.2. STARS4Water project. Horizon Europe Grant Agreement No 101059372.
- Barnett, T. P., Adam, J. C., & Lettenmaier, D. P. (2005). Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature*, 438 (7066), 303-309.
- Barrett, A. P. (2003). National operational hydrologic remote sensing center snow data assimilation system (SNODAS) products at NSIDC.
- Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., ... & Meng, J. (2003). Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *Journal of Geophysical Research: Atmospheres*, 108 (D22).
- Defnet, A., Hasling, W., Condon, L., Johnson, A., Artavanis, G., Triplett, A., Lytle, W., & Maxwell, R. (2024). hf_hydrodata: A Python package for accessing hydrologic simulations and observations across the United States. *Journal of Open Source Software*, 9(99), 6623. <https://doi.org/10.21105/joss.06623>
- Duan, S., Ullrich, P., Risser, M., & Rhoades, A. (2024). Using temporal deep learning models to estimate daily snow water equivalent over the Rocky Mountains. *Water Resources Research*, 60(4), e2023WR035009.
- Duran-Llacer, I., Arumí, J. L., Arriagada, L., Aguayo, M., Rojas, O., González-Rodríguez, L., Rodríguez-López, L., Martínez-Retureta, R., Oyarzún, R., & Singh, S. K. (2022). A new method to map groundwater-dependent ecosystem zones in semi-arid environments: A case study in Chile. *Science of The Total Environment*, 816, 151528. <https://doi.org/10.1016/j.scitotenv.2021.151528>
- Duran-Llacer, I., Gómez-Escalonilla, V., Aliaga-Alvarado, M., Arumí, J. L., Zambrano, F., Rodríguez-López, L., Martínez-Retureta, R., & Martínez-Santos, P. (2025). Approach to mapping groundwater-dependent ecosystems through machine learning in central Chile. *Groundwater for Sustainable Development*, 31, 101526. <https://doi.org/10.1016/j.gsd.2025.101526>
- El Halabi, L., Mital, U., & Dwivedi, D. (2025). Modeling spatial distribution of snow water equivalent using transfer learning across mountainous basins. *Journal of Geophysical Research: Machine Learning and Computation*, 2(2), e2024JH000278.
- FAO. 2026. AQUASTAT Core Database. Food and Agriculture Organization of the United Nations. Database accessed on 2026/06/26.

- Fontrodona-Bach, A., Schaeffli, B., Woods, R., Teuling, A. J., and Larsen, J. R.: NH-SWE: Northern Hemisphere Snow Water Equivalent dataset based on in situ snow depth time series, *Earth Syst. Sci. Data*, 15, 2577–2599, <https://doi.org/10.5194/essd-15-2577-2023>, 2023.
- Fussi, F., Gómez-Escalonilla, V., Rahobisoa, J. J., Ramanantsoa, H. O. A., & Martinez-Santos, P. (2025). An Automated Method to Assess the Suitability of Existing Boreholes for Solar-Based Pumping Systems: An Application to Southern Madagascar. *Sustainability*, 17(3), 1255.
- Gleeson, T., Smith, L., Moosdorf, N., Hartmann, J., Dürr, H. H., Manning, A. H., ... & Jellinek, A. M. (2011). Mapping permeability over the surface of the Earth. *Geophysical Research Letters*, 38(2).
- Kollet, S. J., & Maxwell, R. M. (2006). Integrated surface–groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model. *Advances in Water Resources*, 29(7), 945–958.
- Kratzert, F., et al., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23 (12), 5089–5110
- Kratzert, F., et al., 2023. Caravan - a global community dataset for large-sample hydrology. *Scientific data*, 10 (1), 61. Available from: <http://dx.doi.org/10.1038/s41597-023-01975-w>.
- Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS Opinions: Never train an LSTM on a single basin. *Hydrology and Earth system sciences discussions*, 2024, 1-19.
- Lees, T., et al., 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. *Hydrology and Earth System Sciences*, 25 (10), 5517–5534.
- Lehner, B., Verdin, K., & Jarvis, A. (2008). New global hydrography derived from spaceborne elevation data. *Eos, Transactions American Geophysical Union*, 89(10), 93–94.
- Lehner, B., et al., 2011. High-resolution mapping of the world’s reservoirs and dams for sustainable river-flow management. *Frontiers in Ecology and the Environment*, 9 (9), 494–502.
- Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., Thieme, M. (2019). Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data* 6: 283. doi: <https://doi.org/10.1038/s41597-019-0300-6>
- Maxwell, R. M., Condon, L. E., & Kollet, S. J. (2015). A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3. *Geoscientific model development*, 8(3), 923–937.
- Maxwell, R. M., & Miller, N. L. (2005). Development of a coupled land surface and groundwater model. *Journal of Hydrometeorology*, 6(3), 233–247.
- Muñoz Sabater, J. (2019): ERA5-Land hourly data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: [10.24381/cds.e2161bac](https://doi.org/10.24381/cds.e2161bac) (Accessed on 2025/05/27)
- Musselman, K. N., Addor, N., Vano, J. A., & Molotch, N. P. (2021). Winter melt trends portend widespread declines in snow water resources. *Nature Climate Change*, 11(5), 418–424.K
- O'Neill, M. M., Tijerina, D. T., Condon, L. E., & Maxwell, R. M. (2021). Assessment of the ParFlow–CLM CONUS 1.0 integrated hydrologic model: evaluation of hyper-resolution water balance components across the contiguous United States. *Geoscientific Model Development*, 14(12), 7223–7254.

- Rao, M.M., et al., 2026. Reassessing model complexity in reservoir outflow forecasting: A multi-site, physics-informed benchmark of deep learning and ensemble methods. *Ecological Engineering & Environmental Technology*, 27 (3), 239–252.
- RID, Royal Irrigation Department, <https://app.rid.go.th/reservoir/>, accessed on 2024/11/19
- Rohde, M. M., Albano, C. M., Huggins, X., Klausmeyer, K. R., Morton, C., Sharman, A., Zaveri, E., Saito, L., Freed, Z., Howard, J. K., Job, N., Richter, H., Toderich, K., Rodella, A.-S., Gleeson, T., Huntington, J., Chandanpurkar, H. A., Purdy, A. J., Famiglietti, J. S., Scanlon, B. R., Cuthbert, M. O., Ferguson, G., Taylor, R. G., & Stella, J. C. (2024). Groundwater-dependent ecosystem map exposes global dryland protection needs. *Nature*, 632(8023), 101–107. <https://doi.org/10.1038/s41586-024-07702-8>
- Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y., Devarakonda, R., & Cook, R. B. (2013). *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2*. Oak Ridge National Laboratory (ORNL).
- Thornton, M. M., Thornton, P. E., Wei, Y., Mayer, B. W., Cook, R. B., & Vose, R. S. (2016). *Daymet: monthly climate summaries on a 1-km grid for North America, Version 3*. ORNL DAAC, Oak Ridge, Tennessee, USA.
- Tran, V. N., Nguyen, T. V., Kim, J., and Ivanov, V. Y.: Technical note: Does Multiple Basin Training Strategy Guarantee Superior Machine Learning Performance for Streamflow Predictions in Gaged Basins?, *EGUspHERE* [preprint], <https://doi.org/10.5194/egusphere-2025-769>, 2025
- U.S. Geological Survey. (2024). *National Water Information System: Groundwater levels*. Retrieved [fecha de descarga], from <https://waterdata.usgs.gov/nwis/gw>
- World Urbanization Prospects, United Nations (UN), uri: population.un.org/wup, accessed 2026/06/26
- Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., & Mocko, D. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research: Atmospheres*, 117, D03109. <https://doi.org/10.1029/2011JD016048>
- Xia, Y., Peter-Lidard, C. D., Huang, M., Wei, H., & Ek, M. (2015). Improved NLDAS-2 Noah-simulated hydrometeorological products with an interim run. *Hydrological Processes*, 29(5), 780-792.