



STARS 4 Water

***Future of Europe's freshwater resources:
upscaling basin-wide assessments of climate
change risks and impacts to the EU-level.***

Deliverable 4.4

September 2026

D4.4 FUTURE OF EUROPE'S FRESHWATER RESOURCES

Lead beneficiary	Jülich Research Centre - FZJ
Lead author(s)	Leandro Avila and Stefan Kollet
Work Package	4: Integrated assessment of future water resources
Due date	30 September 2026
Submission date	10 August 2026
Contributors	Héctor Aguilera (IGME), Victor Gomez (UCM), Pedro Martinez (UCM), Helen Baron (UKCEH), Kolbjørn Engeland (NVE), Emiliano Gelati (NVE), Trine Hegdahl (NVE), Stefan Kollet (FZJ), Leandro Avila (FZJ), Peter Salamon (JRC), Stefania Grimaldi (JRC), Berny Bisselink (JRC).

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	20.03.2026	Leandro Avila	Outline
0.6	03.07.2026	Leandro Avila	First draft
0.7	17.07.2026	Trine Jahr Hisdal	Review
0.8	30.08.2026	Kolbjørn Engeland	Review
0.9	07.09.2026	Leandro Avila	Final draft
1.0	18.09.2026	Harm Duel	Accepted

Citation: Avila L., Kollet S., et al., 2026. Future of Europe's freshwater resources: upscaling basin wide assessments of climate change risks and impacts to EU-level. Deliverable D4.4. STARS4Water Project.



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 4.4 (WP4) of the STARS4Water project, which set out to scale up climate change risk and impact assessments from individual river basins to EU level. The central question driving this work is whether data-driven and machine learning models, initially trained at river basin scale, can retain enough predictive skill when applied to much larger and more climatically diverse domains and, if so, whether they can serve as practical, computationally affordable tools for tracking key hydrological variables across Europe.

We evaluated a range of indicators covering both water availability and water quality. Snow Water Equivalent (SWE) and Water Table Depth (WTD) were defined by consulted basin stakeholders as main target variables for deep learning surrogate modelling at a continental scale, given their relevance for seasonal water storage and groundwater recharge.

The report also presents several machine learning models developed within STARS4Water that were successfully transferred to other European river basins. These include a data-driven model for assessing nitrate contamination in the East Anglia Chalk Aquifer, ensemble tree models for reservoir storage forecasting applied to the Seine River Basin, an ensemble LSTM transfer-learning approach for estimating water table depth (WTD) across Europe, and a groundwater storage model implemented in the Seine River Basin.

The results demonstrate the potential of transferable machine learning approaches to support water resource assessment and management across diverse hydrological settings. The clearest finding to emerge 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 losing too much predictive skill, though how much retraining or fine-tuning is needed depends strongly on how similar the target domain is to where the model was originally built. For SWE and WTD, ConvLSTM surrogate models demonstrated genuine spatial transferability, achieving performance levels broadly comparable to the reference physically based models when evaluated against independent in-situ observations. For reservoir storage, direct transfer to basins with different operational characteristics added little over a simple benchmark, while local retraining consistently improved forecast skill. For groundwater quality, the MLMapper tool generalised effectively to a different European aquifer system, and the groundwater storage model transferred successfully to a river basin with different hydroclimatic conditions.

The results also show that globally available reanalysis forcing products, like ERA5-Land, can sustain reasonable surrogate 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. Sensitivity analyses using synthetic what-if scenarios confirmed that the deep learning models respond to hydroclimatic perturbations in physically consistent ways, with temperature emerging as the dominant control on snowpack dynamics and compound drought conditions producing the most spatially extensive groundwater responses across Central Europe. 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 EU-level water resources assessment. All datasets produced in this report are available in the Zenodo portal: <https://zenodo.org/communities/stars4water/records>

Table of contents

Summary	iii
Table of contents	iv
1. Introduction	1
1.1. STARS4Water project	1
1.2. Work Package 4	2
1.3. Content of this report.....	3
2. Predicting daily snow water equivalent	5
2.1. Introduction.....	5
2.2. Study area and data.....	6
2.3. Methods	9
2.4. Results	13
2.5. Conclusions.....	27
3. Estimating daily water table depth	29
3.1. Introduction.....	29
3.2. Study area and data.....	30
3.3. Methods	34
3.4. Results	35
3.5. Conclusions.....	45
4. Predictive mapping of groundwater quality	48
5. Ensemble-tree models for reservoir storage	51
6. Ensemble LSTM transfer learning for Water Table Depth	54
7. Groundwater storage using machine learning models.....	61
8. What-if scenarios to explore future water resources	64
8.1. Introduction.....	64
8.2. Snow Water Equivalent	64
8.3. Water Table Depth	69
9. Conclusions and next steps	74
10. Bibliography	75

1. Introduction

1.1. 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.

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 Rhine, Danube, Messara, Duero, East Anglia, Seine, and Drammen (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.

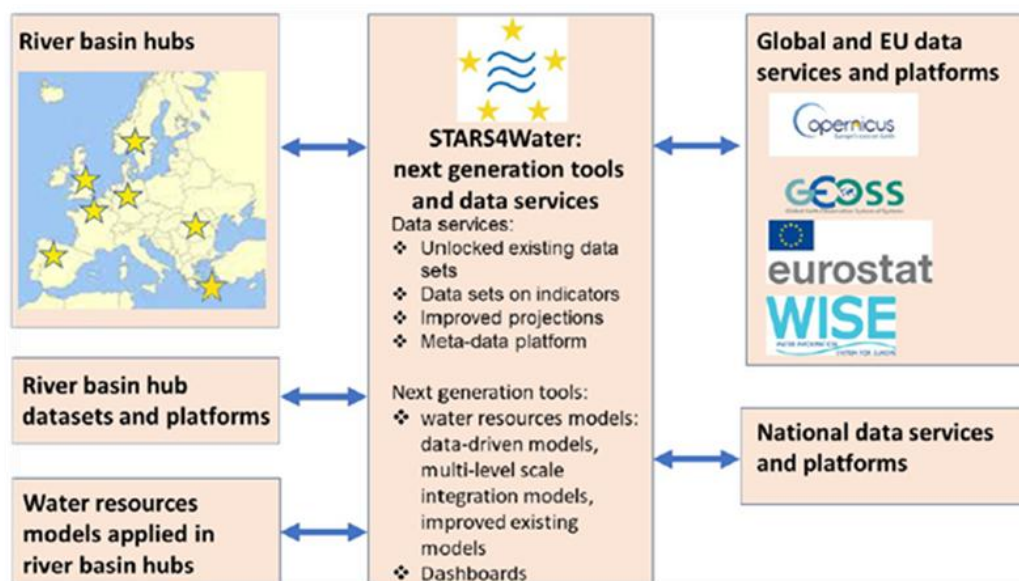


Figure 1.1. Overview of STARS4Water activities within the context of stakeholders and data providers.



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.

1.2. Work Package 4

Work Package 4 (WP4) of the project is dedicated to the integrated assessment of future water resources. The main activities in this work package are to:

- Calibrate and validate the next generation models in the 7 RBHs (Task 4.1).
- Assess and report of water resources availability under risk by climate change in the seven RBHs, by interacting with stakeholders (Task 4.2).
- Provide story maps on climate change and socio-economic developments within the seven RBHs (Task 4.3).
- Upscale basin-wide assessment of climate change risks and impacts to EU-level (Task 4.4).

The interrelation of Work Package 4 in the STARS4Water Framework is displayed in Figure 1.3. Its activities are strongly supported by stakeholders codesign (WP1) and rely on data (WP2) and models (WP3) developed along the process of assessing future water resources for increased preparedness in water resources management at river basin level in Europe (WP5).

D4.4 FUTURE OF EUROPE'S FRESHWATER RESOURCES

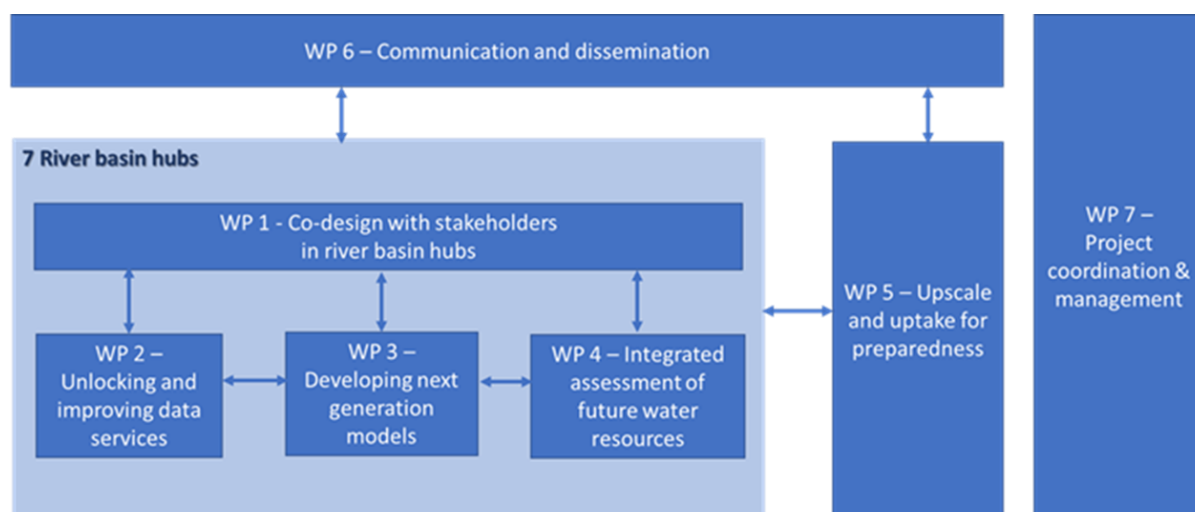


Figure 1.3. Work package structure of the STARS4Water project: the development of the next generation models (WP3) through a co-creation approach (WP1), benefiting from data sources available (WP2) and calibration and validation activities in WP4.

1.3. Content of this report

This report draws on the results of STARS4Water activities in WP3: high-resolution land surface and hydrological model simulations (Avila & Kollet, 2026a), the developed machine learning tools (Baron & Keller, 2026) and the data collections made in WP1 and WP2: observational datasets from national and international sources. For each selected hydrological variable, the report covers:

- Development and training of ConvLSTM surrogate frameworks using physically based model simulations as reference datasets
- Evaluation of spatial transferability from basin-scale training domains to national and continental extents
- Validation against independent in-situ observational records and alternative large-scale model benchmarks
- Application to what-if scenario experiments, including synthetic hydroclimatic perturbations and climate projection scenarios
- Machine learning applications to groundwater quality prediction and reservoir storage forecasting across multiple European river basins.

At a workshop with researchers and stakeholders two hydrological variables were selected to explore the potential of data-driven tools for assessments at EU level: snow water equivalent and groundwater table depth (Avila & Kollet, 2026b).

The report is structured as follows:

- Chapter 2 covers the ConvLSTM surrogate model for daily Snow Water Equivalent prediction, from its development over the Drammen River Basin to its transfer across all of Norway and the European Alps.
- Chapter 3 describes the equivalent framework for Water Table Depth, trained over Central Europe using ParFlowCLM simulations and evaluated across the full pan-European domain against TSMP as an independent benchmark.
- Chapter 4 presents the predictive mapping of nitrate contamination in the East Anglia Chalk aquifer.

- Chapter 5 summarizes results from applying the ensemble tree reservoir storage model to the Seine basin.
- Chapter 6 presents an ensemble LSTM transfer learning for Water Table Depth.
- Chapter 7 introduces machine learning approaches to groundwater storage estimation.
- Chapter 8 presents the analysis of what-if scenario for both ConvLSTM surrogate models presented in Chapter 2 and 3, covering synthetic forcing perturbations and structured climate projection experiments.
- Finally, Chapter 9 provides overall conclusions and outlines next steps toward the operational deployment of these frameworks within STARS4Water.

2. Predicting daily snow water equivalent

2.1. Introduction

Snow Water Equivalent (SWE) is a key variable that characterizes snowpack dynamics and quantifying the amount of water stored in seasonal snow cover (Egli et al., 2009; Taheri & Mohammadian, 2022). It represents the depth of water that would be obtained if the entire snowpack melted completely. In snow-dominated regions seasonal snowpack constitutes a major component of the hydrological cycle, contributing substantially to annual runoff, groundwater recharge, and reservoir inflows (Barnett et al., 2005; Fassnacht et al., 2003; Jenicek & Ledvinka, 2020). Consequently, accurate SWE estimation is essential for hydropower production, flood forecasting, avalanche hazard assessment, and long-term water resources management (Musselman et al., 2021).

In Norway, for example, hydropower represents approximately 90% of national electricity production, making runoff from snowmelt a critical factor for energy planning and reservoir operation (Graabak et al., 2017; Kjaerland, 2007). Reliable SWE information enables operators to anticipate seasonal inflows and optimize reservoir management strategies (Solvang et al., 2014). Similarly, alpine regions across Europe strongly depend on snowpack evolution for water supply, ecosystem functioning, tourism, and hydroelectric production. However, ongoing climate change conditions are affecting snow accumulation and melting processes by increasing winter temperatures, reducing snow persistence, and shifting precipitation from snow to rainfall. These changes increase hydrological uncertainty and amplify the occurrence of extreme events such as rain-on-snow floods (Pall et al., 2019; Mooney & Li, 2021), highlighting the need for more robust and transferable SWE prediction systems.

Traditionally, SWE estimation has relied on in-situ observations, physically based snow models, and satellite remote sensing products. Ground observations provide accurate measurements but are spatially sparse and difficult to maintain in mountainous terrain (Molotch & Bales, 2005). Physically based snow models simulate snow accumulation and ablation processes using meteorological forcing and land surface characteristics, although their performance is often constrained by parameter uncertainty and forcing data quality (Clark et al., 2011; Kavetski et al., 2006). Remote sensing products provide broad spatial coverage but may suffer from coarse spatial resolution, retrieval uncertainty in deep snowpacks, and limitations associated with cloud cover and terrain complexity (Dozier & Painter, 2004; Sheffield et al., 2018). Despite significant progress in data assimilation and hybrid modeling approaches (Margulis et al., 2019; Leisenring & Moradkhani, 2011), generating accurate and high-resolution SWE predictions across heterogeneous regions remains a major challenge.

Recent developments in machine learning (ML) and deep learning (DL) offer new opportunities to improve SWE estimation by exploiting large, multi-source environmental datasets (Bair et al., 2018; Čistý et al., 2024). ML methods can capture highly nonlinear relationships between SWE and environmental controls such as topography, meteorological variables, land cover (Ntokas et al., 2021). Unlike physically based models, deep learning approaches learn directly from data without requiring explicit representation of snow processes, making them particularly suitable for complex mountainous environments (Broxton et al., 2019). Several studies have demonstrated the capability of deep learning models for hydrological prediction, snow reconstruction, and spatial interpolation (Shao et al., 2022; Mital et al., 2022; King et al., 2020).

Among these approaches, convolutional neural networks (CNNs) (Sit et al., 2020) and long short-term memory (LSTM) networks (Kratzert et al., 2018) have shown strong performance in modeling environmental spatiotemporal processes (Le & Behrangi, 2022; Wang et al., 2022). CNNs efficiently extract spatial features from gridded datasets such as topography and meteorological fields (Patakchi Yousefi & Kollet, 2023; Le & Behrangi, 2022), while LSTM architectures capture temporal dependencies associated with snow accumulation and melt evolution (Wang et al., 2022). Hybrid architectures such as ConvLSTM combine these capabilities by simultaneously learning spatial and temporal relationships, making them especially suitable for SWE prediction tasks. Previous studies have shown that ConvLSTM models outperform standalone CNN and LSTM approaches in several hydrometeorological applications (Wu et al., 2020; Barzegar et al., 2020). Furthermore, surrogate deep learning models have recently emerged as efficient alternatives for emulating computationally expensive physical models while preserving their large-scale dynamics (Sun et al., 2023).

Building upon these advances, this study investigates the use of a ConvLSTM-based surrogate model to predict daily SWE over multiple regions in Europe. A key component of this task is the assessment of spatial transferability and model generalization. This approach aims to explore the potential of transferable deep learning surrogates for large-scale snow hydrology applications, reducing dependence on extensive local training and facilitating deployment in data-scarce mountainous regions. The model was initially trained over the Drammen River basin in southern Norway and subsequently transferred to larger and climatically distinct regions, including all of Norway and the European Alps. The primary objective is to evaluate whether a deep learning model trained in a relatively limited domain can generalize its learned spatiotemporal snow dynamics to broader geographical regions characterized by different topographic, climatic, and snow regimes. Instead of relying directly on sparse observations, the model is trained using the SeNorge SWE dataset (Saloranta, 2012), which provides consistent daily snow information at 1 km spatial resolution across Norway. Meteorological forcing variables, topographic descriptors, and temporal indicators are used as model inputs to predict SWE up to seven days ahead.

2.2. Study area and data

Study areas

Figure 2.1 depicts the location of the three study areas: the Drammen river basin, Norway and the Alps. The Drammen river basin is located in southeastern Norway and encompasses a catchment area of approximately 17,000 km². The basin's hydrological regime is heavily influenced by seasonal snow accumulation and melt, particularly in its mountainous areas, where snowpack acts as a natural reservoir. During winter, snow stores up to 60–70% of the basin's annual precipitation, which is gradually released during spring and summer melt periods (Lussana et al., 2018). This delayed release sustains baseflow in rivers, and ensures consistent water supply for hydropower generation, which accounts for over 90% of Norway's electricity production (Aasen et al., 2010). Effective management of snow-derived water is thus important for maintaining energy security and operational stability in hydropower infrastructure. Snowmelt also plays a pivotal role in flood risk mitigation. Rapid spring snowmelt, compounded by rainfall, has historically triggered destructive floods, such as the 2007 and 2013 events that caused extensive damage to downstream communities (Hailemariam & Alfredsen, 2023). Reservoir operators rely on real-time SWE data from models like SeNorge and in situ measurements to forecast melt volumes and regulate storage capacity, balancing flood prevention with hydropower needs (Saloranta, 2012).

The Alps extend across several European countries, including France, Switzerland, Italy, Austria, and Germany, and is often named the “water tower of Europe” as they supply a substantial fraction of the streamflow in several large European rivers such as the Rhine, Po, and Danube. The hydrological regime of these basins is heavily influenced by seasonal snow accumulation and melting, particularly at elevations above 1,500 m, where the snowpack acts as a natural reservoir. During winter, snow stores up to 50–70% of the annual precipitation in high-altitude catchments, which is gradually released during spring and summer melt periods (Barnett et al., 2005). This delayed release sustains baseflow in rivers and ensures consistent water supply for hydropower generation, agriculture, and municipal use across much of continental Europe. Effective management of snow-derived water is thus important for maintaining energy security and operational stability in hydropower infrastructure, which is particularly critical in countries like Switzerland and Austria, where hydropower accounts for over 50% of national electricity production (Schuler et al., 2006).

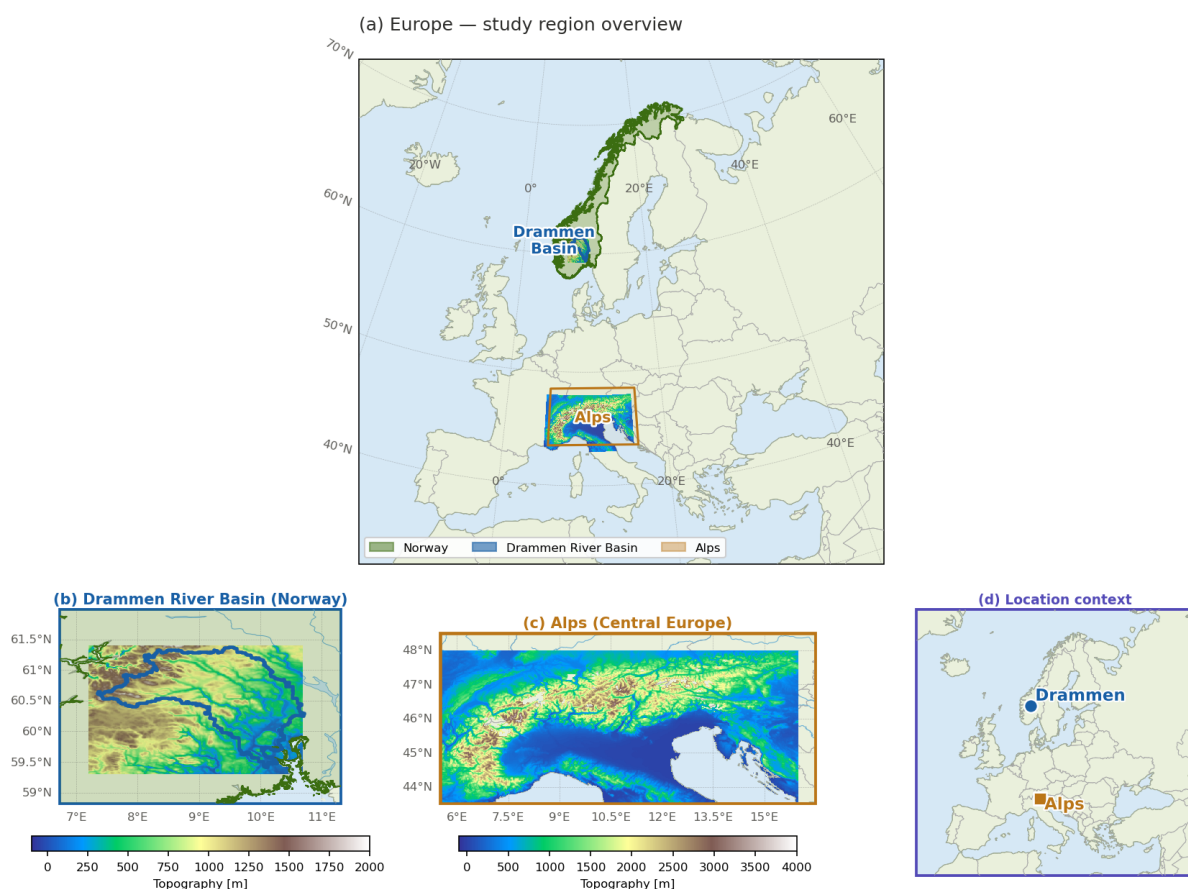


Figure 2.1. Study areas considered to apply a ConvLSTM model to predict daily snow water equivalent time series.

SeNorge data set and snow model

The SeNorge dataset provides a high-resolution (1 km x 1 km) observational grid of multiple climatological variables in Norway, including daily precipitation and temperature fields from 1957-present (Lussana et al., 2018; Lussana et al., 2019). Developed to address the challenges posed by Norway's complex topography and sparse observational coverage in remote areas, the dataset is derived by using advanced spatial interpolation techniques that combine station observations with physiographic predictors (Lussana et al., 2018). The SeNorge observational grid dataset does not include information about snow cover. Instead, daily maps of snow conditions are produced by the Norwegian Water Resources and Energy Directorate (NVE) with the SeNorge snow model since 2004 (Saloranta, 2012).

The SeNorge snow model, which relies on precipitation and temperature, estimates, among others, snow water equivalent (SWE), snow depth (SD) and snow bulk density (ρ) (Saloranta, 2012; Saloranta, 2016). Previous studies evaluated the performance of the SeNorge model against in-situ observations. The authors suggested that the model structure is suited for describing snow conditions at ungauged sites and for changing climate conditions (Saloranta, 2012; Skaugen et al., 2018).

ERA5-Land reanalysis

The ERA5-Land reanalysis dataset provides temporally consistent representation of key land surface variables spanning several decades (Muñoz-Sabater et al., 2019). Developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), this dataset is produced using the ECMWF land surface model. The simulations are forced by meteorological inputs from the ERA5 climate reanalysis, which undergo dynamical downscaling and elevation-based corrections to enhance the accuracy of near-surface thermodynamic variables, including temperature and humidity (Muñoz-Sabater et al., 2019). With a horizontal resolution of approximately 9 km, ERA5-Land offers a substantial improvement over the 31 km resolution of its parent ERA5 dataset (Johnson et al., 2019). This finer spatial resolution renders the dataset particularly useful for applications demanding detailed land surface information, such as hydrological modeling, water resource management, agricultural planning, land use analysis, and environmental monitoring (Muñoz-Sabater et al., 2019). Covering the period from January 1950 to near real-time (with a five-day latency), ERA5-Land is publicly available through the Copernicus Climate Data Store. For this study, we used 2-meter air temperature and daily precipitation data from 2020–2024 (retrieved at hourly resolution) to validate the proposed machine learning framework. The dataset was interpolated to the same spatial resolution as the SeNorge_2018 dataset using bicubic interpolation methods. In addition, a bias correction approach based on the quantile-quantile map was executed for both variables (precipitation and temperature) to approximate the spatial variance observed by the SeNorge_2018 dataset (Teutschbein & Seibert, 2012; Xie et al., 2022).

LISFLOOD EFAS

LISFLOOD simulations derived from the European Flood Awareness System (EFAS) were used to validate the ConvLSTM model over the Alps regions. Those simulations were produced by JRC and were also used and presented by Avila and Kollet (2025) to perform the multi-scale model integration by comparing multiple large-scale models. This LISFLOOD model setup was forced with EMO-1 (European Meteorological Observations) gridded sub-daily dataset at a 1x1 arcminute resolution (~1.5 km at EFAS latitudes) (Gomes et al., 2020) across the European Flood Awareness System (EFAS) domain. For the latest version data is available from 1992-01-01 up until near real time. The output variables include river discharge in the last 24 and 6 hours, snow depth water equivalent, and volumetric soil moisture. We will hereby refer to this model configuration as LISFLOOD-Europe.

In-situ observations: the Northern Hemisphere Snow Water Equivalent (NH-SWE)

We used the NH-SWE dataset to compare and validate the ConvLSTM model against snow water equivalent values derived from in-situ observations. The NH-SWE dataset is derived from existing, freely available in situ snow depth measurements, which are compiled from multiple national and international sources, such as the Norwegian Meteorological Institute (Fontrodona-Bach et al., 2023). For this dataset, the authors converted the corresponding snow depth to snow water equivalent (SWE) values based on the Δ SNOW model (Winkler et al., 2021) by simulating snowpack processes such as new snow accumulation, dry compaction, and melt, using regionally calibrated parameters for fresh snow density and maximum snow density. The dataset spans 1950 to 2022, with station time

series ranging from 1 to 73 years in length. Each station provides daily SWE estimates, modelled snow density, and quality control flags, though the original snow depth data must be obtained separately from the original sources. This in situ-based approach offers a valuable complement to remote sensing and reanalysis products, particularly for capturing point-scale snow dynamics in mountainous and data-sparse regions. Although these are not direct SWE measurements, they represent a good approximation and were used for evaluation of model results.

2.3. Methods

This study applied an autoregressive ConvLSTM framework to generate snow water equivalent (SWE) maps with a 7-day prediction horizon. The model uses a combination of dynamic and static inputs, including 60-day sequences of historical precipitation, temperature, and SWE; 7-day projections of future precipitation and temperature; topographic variables such as elevation and geographic coordinates; and temporal information represented by the day of the year. Given that SWE is included as an input variable, and the model is designed to operate autoregressively, one of the main objectives was to enable the use of previous predictions as inputs during the simulation process. To mitigate error accumulation associated with this approach, a scheduled sampling strategy was implemented during training, progressively replacing observed SWE inputs with model-generated predictions.

The model was initially trained and validated in the Drammen River Basin, using daily gridded SWE data at a spatial resolution of $1 \text{ km} \times 1 \text{ km}$ produced by the SeNorge snow model. Subsequently, the model was transferred to larger regions and different domains, such as Norway and the Alps. To build the ConvLSTM framework, a 60-day historical input window was selected based on a sensitivity analysis, aiming to maximize the inclusion of antecedent meteorological conditions, including information from previous seasons that may influence snow accumulation and melt processes while maintaining a reasonable computational cost.

The training period spans 2010–2017, while validation and testing were conducted over 2018–2020 and 2020–2024, respectively. To further evaluate model robustness, potential overfitting, and transferability, the trained framework was also forced with precipitation and temperature data from the ERA5-Land reanalysis dataset, which provides globally available meteorological information. A summary of the datasets and model configuration used in this study is presented in Table 2.1.

ConvLSTM cell

ConvLSTM is a hybrid deep learning approach that merges CNN and LSTM networks, being an effective modelling tool for processing spatiotemporal sequence data (Yao et al., 2023). Compared to LSTM, the ConvLSTM cell replaces traditional matrix multiplication operations with convolutional filters for processing both input data and state transitions. This allows the model to simultaneously capture spatial patterns through convolutional operations while maintaining temporal dependencies via LSTM gating structures. By taking advantage of the memory gate of the LSTM, the ConvLSTM can determine the future state of a cell in the grid from the inputs from past states of neighbouring cells (Moishin et al., 2021; Shi et al., 2015). Like the LSTM cell, ConvLSTM includes input gates (I_t), forget gates (F_t), output gates (O_t), hidden states (H_t), and cell states (C_t). Each of these states and gates can be calculated as presented in Figure 2.2, where W represents weight factors and b bias factors.

Autoregressive ConvLSTM framework

The proposed modeling framework used to predict daily SWE was designed based on ConvLSTM units that preserve two-dimensional spatial information during both learning and prediction phases (Figure 2.3). Unlike pipeline architectures that process spatial and temporal components separately, the ConvLSTM maintains spatial correlations within hidden state representations while simulating temporal progression, offering significant advantages for grid-based environmental modeling (Khorram and Jehbez, 2023).

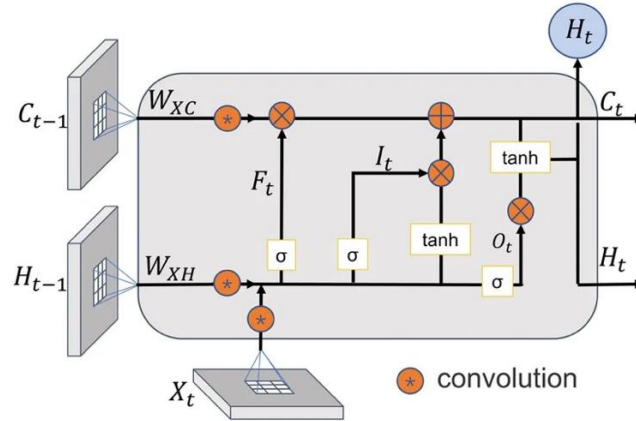


Figure 2.2. Inner structure of ConvLSTM (Shi et al., 2022).

$i_t = \sigma(W_i * [H_{t-1}, X_t] + b_i)$	Equation 1
$f_t = \sigma(W_f * [H_{t-1}, X_t] + b_f)$	Equation 2
$o_t = \sigma(W_o * [H_{t-1}, X_t] + b_o)$	Equation 3
$C = f_t \circ C_{t-1} + i_t \circ g_t$	Equation 4
$H_t = o_t \circ \tanh(C_t)$	Equation 5

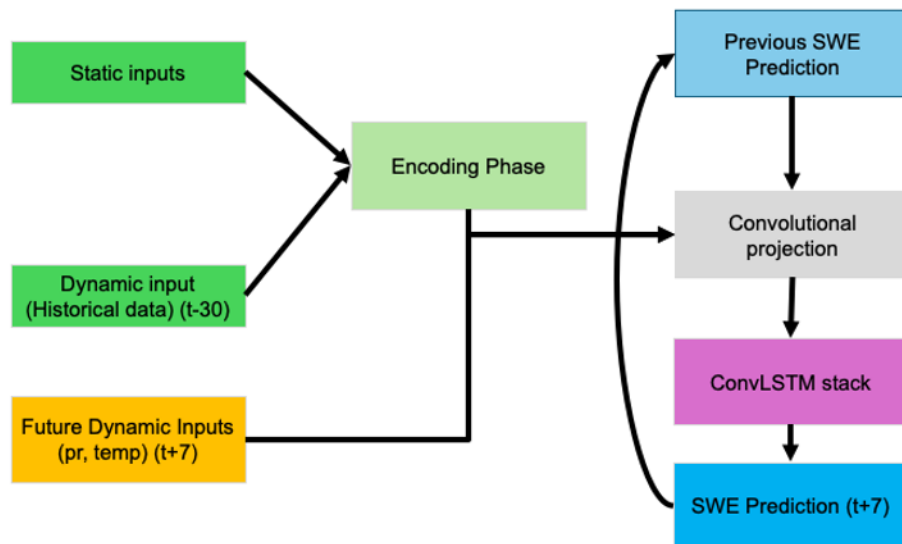


Figure 2.3. Flowchart of the proposed ConvLSTM framework used to emulate the SeNorge model and predict daily SWE.

The operational pipeline consists of two primary stages: historical context integration and sequential prediction generation. In the initial phase, the system processes 60 consecutive days of input sequences containing SWE measurements, precipitation and temperature records, and cyclical temporal indicators. Each temporal instance is first transformed through convolutional layers before entering stacked ConvLSTM modules. These recurrent components maintain evolving hidden and cell states as three-dimensional tensors, capturing both geographical distributions and chronological patterns of snowpack development. Parallel to this dynamic processing, static attributes including elevation and geographical coordinates are encoded through dedicated convolutional networks. During the sequential prediction stage, the model generates daily SWE estimates across a seven-day horizon. For each forecast step, the model synthesizes multiple information streams: previously predicted SWE values, anticipated meteorological conditions, and temporal positioning data. These combined inputs are merged with the static landscape encodings and processed through the ConvLSTM layers, which advance their internal states temporally. The resulting hidden representations are converted to final SWE predictions through convolutional output layers. A key training component involves progressive scheduled sampling (Bengio et al., 2015), where the model gradually shifts from utilizing true SWE values to employing its own prior estimates as subsequent inputs. This approach allows to reduce error propagation during extended forecasting sequences, improving the autoregressive modelling approach. Table 2.1 depicts the main parameters and settings defined during the training process for the proposed ConvLSTM framework, whereas Table 2.2. presents the training parameters used to train the ConvLSTM model.

Table 2.1. Description of datasets and model configuration for the ConvLSTM SWE prediction framework.

Component	Description
Primary Dataset	SeNorge observational dataset and snow model
Alternative Dataset	ERA5-Land reanalysis (bias-corrected)
Input Variables	• Past meteorological conditions (60-day history): daily precipitation (mm/day), mean daily temperature (°C), Snow Water Equivalent (mm) - (SeNorge snow model and ConvLSTM outputs). • Future meteorological conditions (7-day horizon): daily precipitation (mm/day), mean daily temperature (°C). • Static features: topography and latitude. • Temporal indicators: day of year (sin and cos).
Target Variable	Snow Water Equivalent (mm) - SeNorge snow model
Training Period	2010–2017
Validation Period	2018–2020
Testing Period	2020–2024

Table 2.2. Training parameters of the ConvLSTM framework.

Parameter	Value
ConvLSTM layers	2
Hidden dimension size	64
Kernel size	3x3
Optimizer type	Adam
Batch size	2
Training epochs	50
Learning rate	0.0005

Bias correction

The original ERA5-Land reanalysis dataset was first interpolated to the target 1 km spatial resolution, and bias correction was then applied using the high-resolution SeNorge observational dataset. We employed an empirical cumulative distribution function (CDF) matching technique, also known as quantile-quantile (Q-Q) mapping (Kim et al., 2022). This non-parametric method is well suited for correcting distributional biases in climate variables, as it does not assume any specific underlying probability distribution for the data. The bias correction was applied to each grid cell independently to account for local topographic and climatic effects. For a given variable and a specific grid cell, the empirical CDF of the ERA5-Land data (F_{ERA5}) was matched to the empirical CDF of the reference SeNorge data ($F_{SeNorge}$) over a common calibration period. For any raw ERA5-Land value (x_{raw}), the bias-corrected value (x_{bc}) was computed as:

$x_{bc} = F_{SeNorge}^{-1}(F_{ERA5}(x_{raw}))$	Equation 6
--	------------

This transformation ensures that the bias-corrected ERA5-Land data reproduces the same statistical distribution, including mean, variance, and extreme quantiles as the SeNorge dataset for the training period. For precipitation, a threshold was applied to correct the frequency of wet days (e.g., days with precipitation > 0.1 mm/day). After calibration, the same quantile mapping parameters were applied to the independent evaluation period (2021–2024).

Evaluation metrics

The ConvLSTM model was trained following a Bias-Aware Loss function ($\mathcal{L}_{BiasAware}$) that augments the mean squared error (MSE) with a squared bias penalty term, enabling the model to minimize both random and systematic errors (Equation 7). In this case, γ represents the weight for the bias penalty term. The use of this objective function promotes more accurate and unbiased predictions across space and time. In addition, the performance of the proposed model was validated using the Kling-Gupta Efficiency (KGE) metric and its components. The KGE is calculated following Equation 8, where r is the correlation component (Equation 9), α is the variability ratio component (Equation 10) and β is the bias ratio component (Equation 11). We also considered the Nash–Sutcliffe model efficiency coefficient (NSE) to evaluate how well the model fit the peak values of SWE presented across the years (McCuen et al, 2006). For all cases, o denotes the observations and s denotes the simulations.

$\mathcal{L}_{BiasAware} = \frac{1}{N} \sum_{i=1}^N (s_i - o_i)^2 + \gamma \left(\frac{1}{N} \sum_{i=1}^N (s_i - o_i)^2 \right)$	Equation 7
$KGE = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}$	Equation 8
$r = \frac{conv(o, s)}{\sigma_o \sigma_s}$	Equation 9
$\alpha = \frac{\sigma_s}{\sigma_o}$	Equation 10
$\beta = \frac{\mu_s}{\mu_o}$	Equation 11
$NSE = 1 - \frac{\sum_{i=1}^n (o_i - s_i)^2}{\sum_{i=1}^n (o_i - \bar{o})^2}$	Equation 12

2.4. Results

Drammen river basin

We first performed an exploratory analysis to evaluate how well the inputs and target variables are correlated. Figure 2.4 presents an exploratory correlation analysis among long-term mean climatic and geographic variables to guide feature selection for predicting SWE using the proposed modeling framework. The analysis reveals strong positive correlations between SWE and both topography (0.74) and latitude (0.68), as well as a strong negative correlation with temperature (-0.73), indicating these variables are highly relevant for SWE estimation. Precipitation also shows a moderate positive correlation with SWE (0.60), suggesting it contributes additional predictive value. These relationships support the inclusion of both static and dynamic inputs in a hybrid deep learning model.

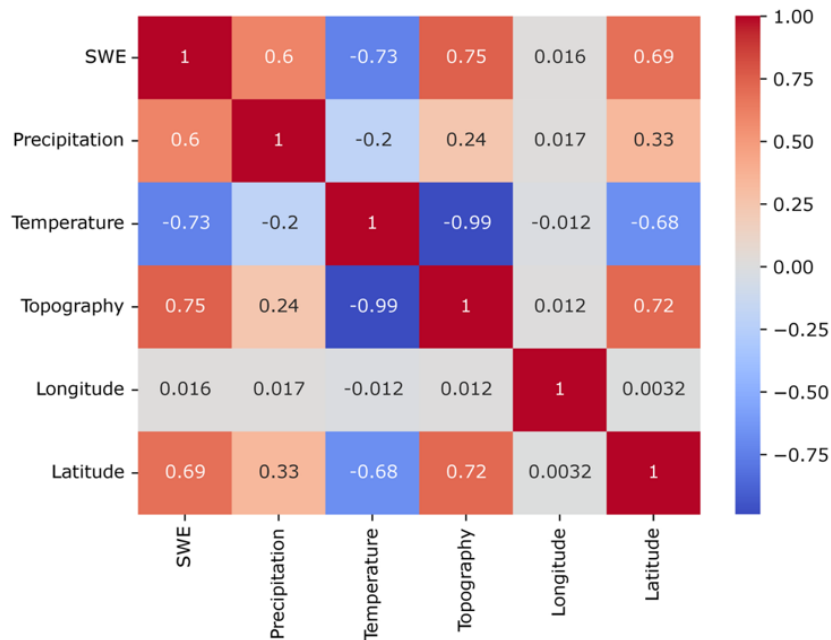


Figure 2.4. Correlation matrix based on Pearson correlation coefficient between different climatic and geographic variables considered for training the ConvLSTM model in the Drammen river basin.

Figure 2.5 compares observed and simulated 7-day ahead SWE time series at six randomly selected locations within the study domain, alongside an aggregated time series that illustrates the overall model performance using two distinct climate forcing datasets (SeNorge and bias-corrected ERA5). The aggregated plot shows that both simulations successfully reproduce the seasonal pattern of SWE accumulation and melt, though the SeNorge-driven model generally achieves closer agreement with observed values. Over multiple seasons, the SeNorge-based predictions more accurately capture both the magnitude and timing of peak SWE compared to those forced with ERA5-Land data.

Considerable spatial variation in model performance emerges across individual sites. Overall, ConvLSTM-ERA5 (blue) systematically underestimates peak SWE values relative to the SeNorge reference, with this underestimation being most pronounced at Points C, E, and F, where the ERA5-driven simulation falls well below the SeNorge curve during peak accumulation periods. In contrast, ConvLSTM-SeNorge (red) tends to slightly overestimate SeNorge values at most locations (A, B, D, and F), closely tracking but consistently exceeding the reference during peak SWE. This pattern is corroborated by the KGE values reported for each station, which are consistently lower for the ERA5-forced configuration than for the SeNorge-forced one, confirming the comparatively reduced skill of the coarser global reanalysis forcing. This shortcoming of ERA5 likely results from its coarser spatial resolution, which struggles to resolve localized orographic precipitation and microclimatic effects that are critical in mountainous areas. Overall, the figure underscores the crucial role of accurate input forcing data in SWE modeling and highlights the added benefit of high-resolution regional datasets like SeNorge, while also reinforcing the spatial heterogeneity of snow processes across Norway and the need for location-specific assessments.

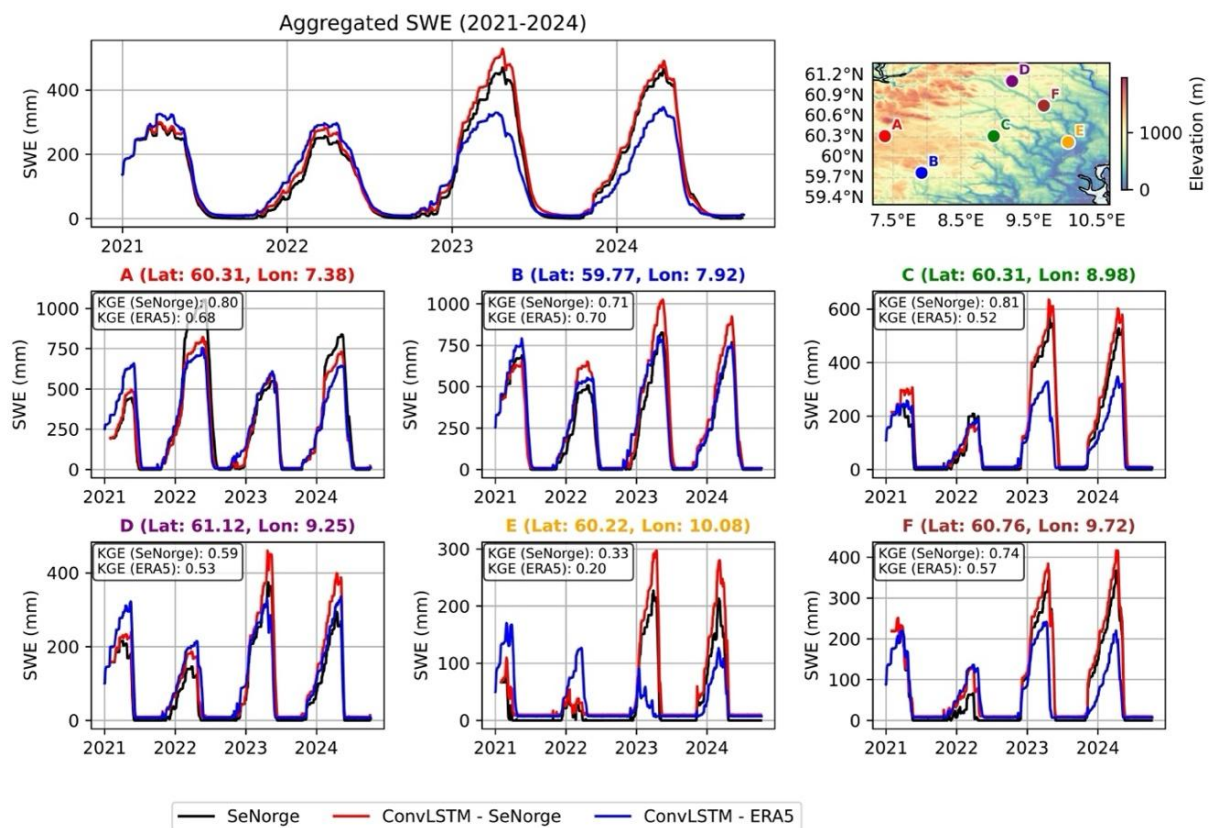


Figure 2.5. Comparison of observed and simulated daily SWE time series at six random selected grid points in the Drammen River basin. The top-left panel shows the basin-aggregated SWE, while the other six panels display SWE at individual grid points (A–F).

Based on the results presented in Figure 2.5, we evaluated the ConvLSTM model performance using the KGE and its individual components (correlation, variability ratio α and bias ratio β), for all grid cells in the study area, including the river basin drainage area, as shown in Figure 2.6. As expected, the model demonstrates better performance when forced with SeNorge data, exhibiting higher median KGE values (~ 0.75) with notably low interquartile ranges, in particular for the correlation and the ratio α . While ERA5-forced simulations result in lower performance in comparison with the simulations forced by SeNorge data, the results are reasonable. However, ERA5 demonstrates significant potential for global applications. As a globally consistent reanalysis, it enables transferability of the modeling framework to other snow-dominated regions where high-resolution inputs are often unavailable. KGE values >0.6 with ERA5 forcing confirm the model's ability to extract meaningful snow dynamics from coarser meteorological inputs.

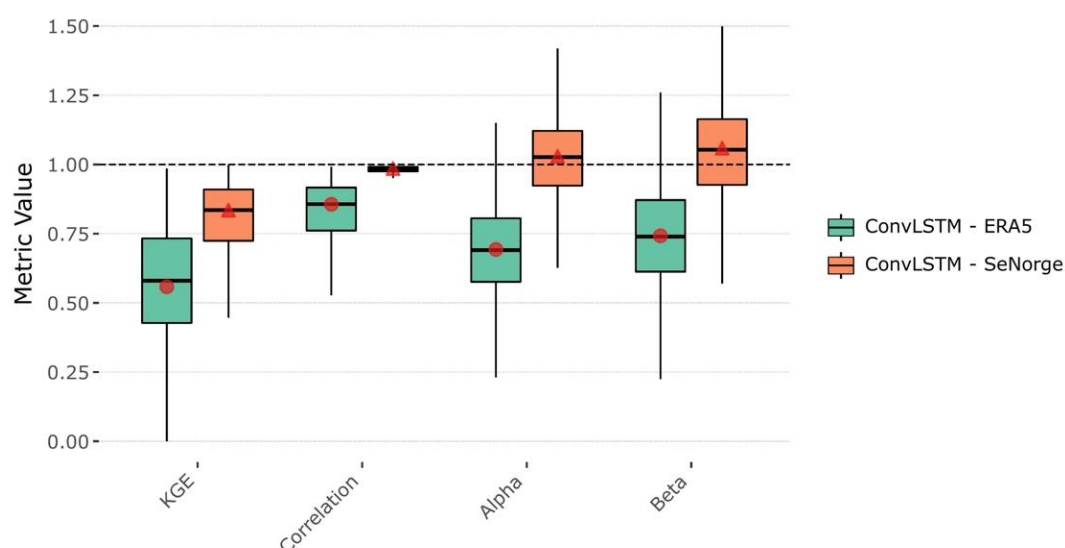


Figure 2.6. Comparison of model performance metrics for SWE simulations using ERA5 and SeNorge datasets. Boxplots show the distribution of KGE, correlation, α , and β metrics across test locations. Red markers indicate the median values for each metric and dataset. The dashed line at 1.0 marks ideal model performance.

Figure 2.7 shows scatter plots that compare daily SWE changes derived from the ConvLSTM model, forced with SeNorge and ERA5 data, against those from the SeNorge snow model. Here, daily SWE changes are defined as the difference in SWE between consecutive days, enabling a direct assessment of the model's ability to simulate both snow accumulation and ablation. The plot is divided into four quadrants, indicating prediction accuracy: the top-right quadrant represents correct accumulation predictions, the bottom-left quadrant correct melt predictions, the top-left quadrant false melt predictions (where accumulation was observed), and the bottom-right quadrant false accumulation predictions (where melt was observed). Overall, both model configurations capture the direction and magnitude of most SWE changes effectively, with most points falling in the correct-prediction quadrants. However, the SeNorge-forced model performs better at reproducing extreme accumulation events and exhibits fewer points in the false-prediction quadrants, likely due to the higher resolution and regional calibration of its input data.

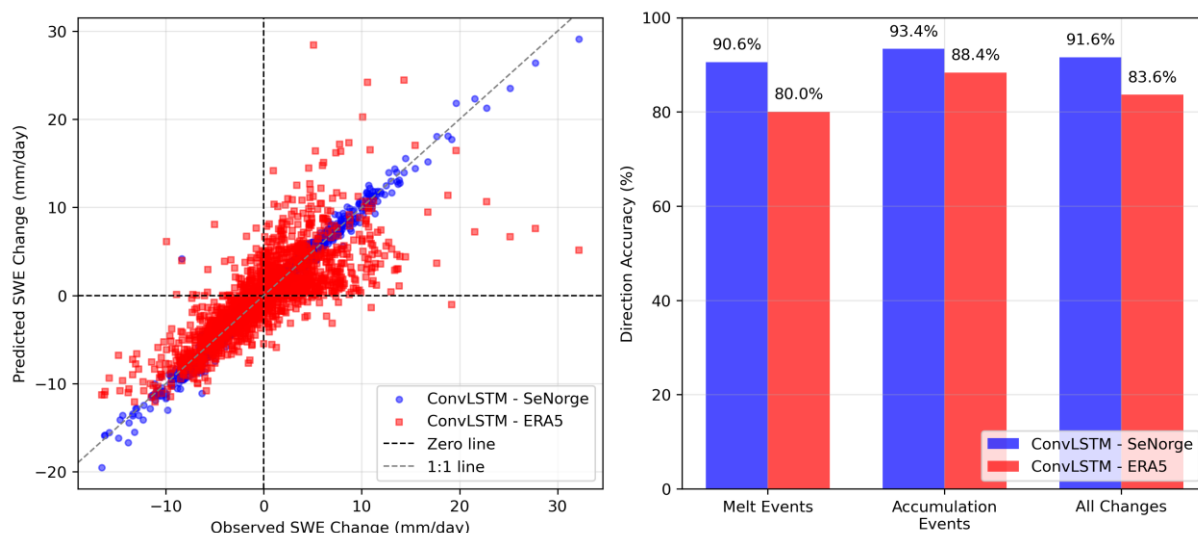


Figure 2.7. Daily SWE change prediction accuracy for ConvLSTM models using SeNorge and ERA5 forcing.

The ability of the ConvLSTM model to reproduce daily SWE was also evaluated against in-situ observations from the NH-SWE dataset using the Pearson correlation coefficient (Figure 2.8) and KGE (Figure 2.9). This dataset includes daily SWE time series spanning 1950–2022 with 132 stations that are located both inside and outside the Drammen river basin. For comparison, each station was matched to the nearest grid pixel that best aligns in terms of elevation and mean annual maximum SWE values between observed and simulated values.

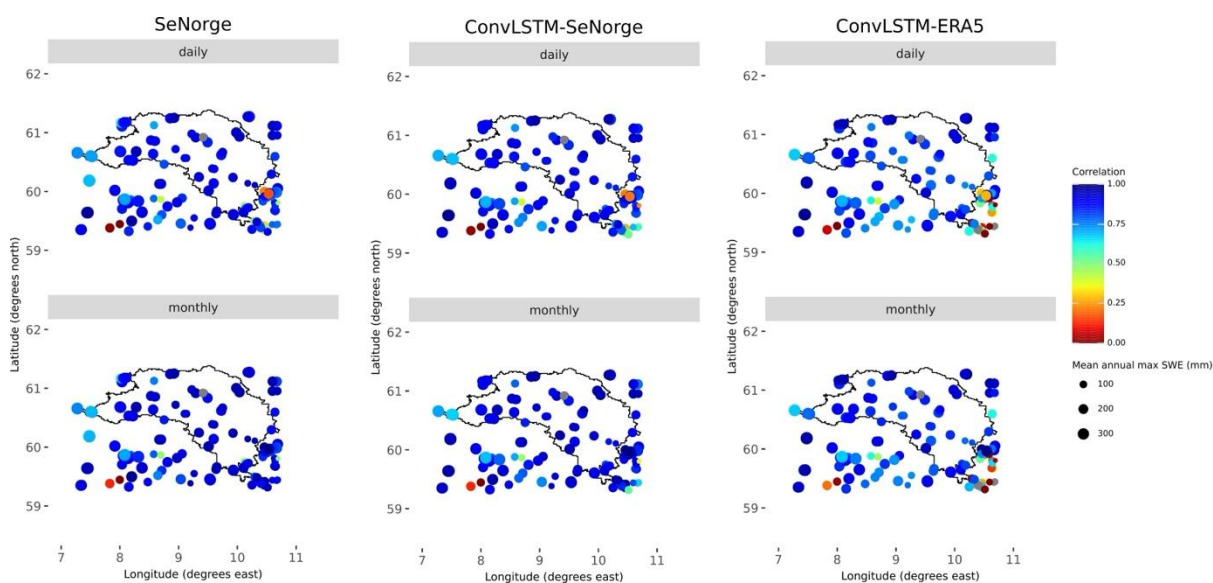


Figure 2.8. Pearson correlation comparison between simulated SWE derived from ConvLSTM and ERA5, and in-situ observations in the Drammen river basin.

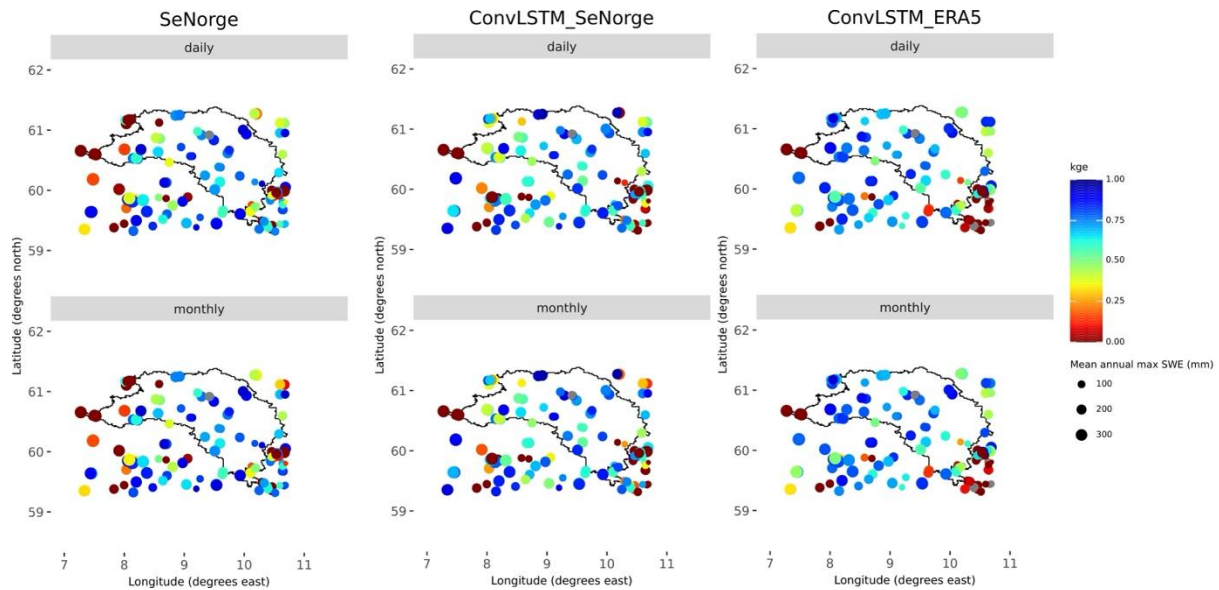


Figure 2.9. KGE comparison between simulated SWE derived from ConvLSTM and ERA5, and in-situ observations in the Drammen river basin.

Figures 2.8 and 2.9 present results from the reference SeNorge dataset alongside ConvLSTM outputs driven by both SeNorge and ERA5 inputs. Overall, the KGE and correlation metrics indicate that the simulated SWE captures the seasonal cycle well, successfully representing the main winter accumulation and spring melt phases at most stations across all three datasets. It is worth mentioning that clear deviations appear in certain locations, especially at high elevations or within topographically complex valleys. At low and mid-elevation stations, however, simulations and observations show reasonably good agreement, suggesting that the ConvLSTM model generalizes effectively beyond the specific characteristics of the SeNorge training region. It should also be noted that, when comparing model results with point measurements, snow depth and SWE can vary substantially within a single grid cell due to wind-driven snow redistribution. As a result, point observation may not accurately reflect the average conditions of the corresponding grid cell simulated by either the ConvLSTM model or the SeNorge reference dataset. This scale mismatch is particularly pronounced in mountainous or coastal areas, where wind effects strongly influence snow cover distribution.

Finally, Figure 2.10 provides a quantitative synthesis of model performance against in-situ observations. In this case, we present boxplots of the KGE values and their components. When forced with SeNorge data, the ConvLSTM model achieves KGE values exceeding 0.70 with low bias (β close to 1) and high correlation ($r > 0.8$). This indicates that the surrogate not only can well emulate SeNorge's behavior but also retains meaningful skill against ground snow measurements. More significantly, the ERA5-driven model also achieves acceptable performance against in-situ observations, with KGE values above 0.6 at a substantial fraction of stations. Although a systematic underestimation persists in the high SWE range, the model successfully captures the overall accumulation and melt dynamics across most of the domain. This result is particularly encouraging, as it demonstrates that a surrogate deep learning model trained on a high-resolution regional product (SeNorge) can be transferred to a global forcing dataset (ERA5) without completely losing predictive skill.

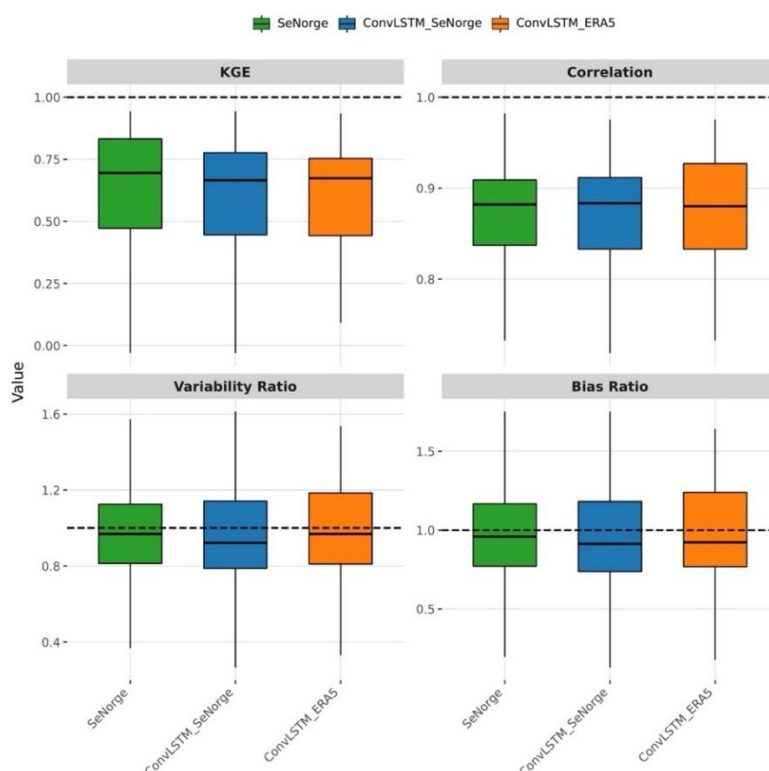


Figure 2.10. Summary performance of the comparison between simulated SWE derived from ConvLSTM and ERA5, and in-situ observations in the study area of the Drammen river basin.

Norway

The ConvLSTM model trained in the Drammen river basin region was subsequently used to predict daily SWE across all of Norway, using the same input datasets (SeNorge and ERA5). It is worth noting that in these cases the model required neither retraining nor any modification of its weights, given that the ConvLSTM architecture operates through local convolutions whose parameters are intrinsically independent of the spatial resolution of the domain. The convolutional filters, once trained, learn relative spatial patterns (temperature gradients, topographic snow distribution, elevation influence) that are equally applicable to any input grid. In practice, this means that the spatial dimensions $H \times W$ of the input tensor are simply replaced by those corresponding to the national Norwegian grid, while the hidden states of the network are dynamically reinitialized to accommodate the new domain. This translational equivariance property of convolutional operations allows the model to be extended from a single basin to a national domain without any structural modification of the architecture. Nevertheless, if the target region presents substantially different climatic or physiographic conditions, such as distinct precipitation regimes, vegetation cover, or orographic complexity, a fine-tuning stage can be applied, updating the model weights on a limited set of local observations. This process allows the model to adapt its learned representations to the specific conditions of the new domain while preserving the general spatiotemporal patterns acquired during the original training, resulting in improved predictive performance with a significantly reduced computational cost compared to training from scratch.

Figure 2.11 presents a spatially distributed evaluation of the ConvLSTM model performance across Norway through four complementary metrics: Percentage Bias (Pbias %), Pearson Correlation (r), coefficient of determination (R^2), and KGE, computed pixel-wise using the SeNorge snow model as reference dataset. The correlation map reveals the most spatially consistent performance, with values predominantly ranging between 0.8 and 1.0 across most of the Norwegian domain, confirming that

the model reliably reproduces the temporal phasing of snow accumulation and melt cycles regardless of geographical location. R^2 values follow a similar spatial pattern, with strong performance concentrated in the interior and southern regions of Norway, and progressive degradation toward coastal and high-elevation zones in the west and north. The Pbias map reveals a systematic overestimation in northern and coastal areas, likely attributable to the mismatch between the continental snow accumulation regime represented in the training data and the more transient, rain-influenced snowpack dynamics characteristic of these regions. The KGE map synthesizes these behaviors, showing values above 0.6 across most of the domain, with localized deterioration in coastal areas of western Norway and in the northernmost regions, consistent with the known limitations of the model in resolving fine-scale snow dynamics driven by maritime climate influences and strong orographic gradients.

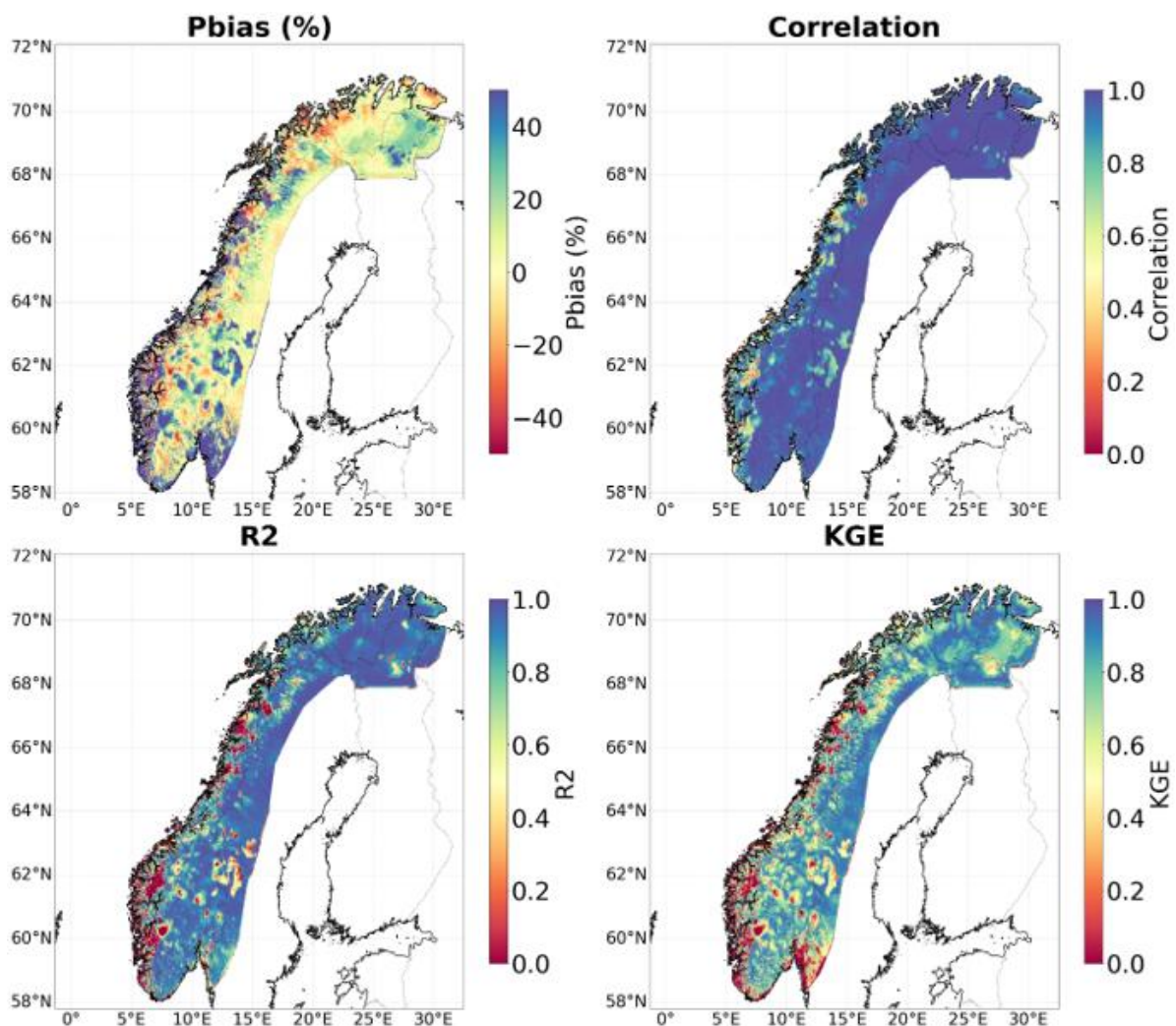


Figure 2.11. Spatial distribution of ConvLSTM model performance metrics evaluated against the SeNorge snow model reference dataset across Norway.

Figure 2.12 summarizes the overall performance of the ConvLSTM model across Norway through boxplots of five key evaluation metrics: NSE, KGE, Pearson Correlation (R), variability ratio (α), and bias ratio (β). The model achieves consistently strong results, with median NSE values approaching 0.9, indicating that the model explains most of the variance in SWE dynamics at the national scale. KGE median values of approximately 0.8, while showing a wider interquartile range and occasional

outliers extending below 0.4, confirm that the model maintains reliable overall performance despite localized degradation in climatically challenging areas. The correlation component (r) exhibits the tightest distribution, with median values close to 0.95 and minimal spread, reflecting the model's robust ability to reproduce the temporal phasing of snow accumulation and melt cycles across diverse Norwegian landscapes. The α and β components cluster tightly around 1.0, suggesting limited systematic bias in both variability and mean SWE magnitude at the domain scale. Collectively, these results demonstrate that the ConvLSTM surrogate model, originally trained over the Drammen River Basin, generalizes effectively to the full Norwegian domain, preserving high predictive skill across the primary dimensions of hydrological model performance.

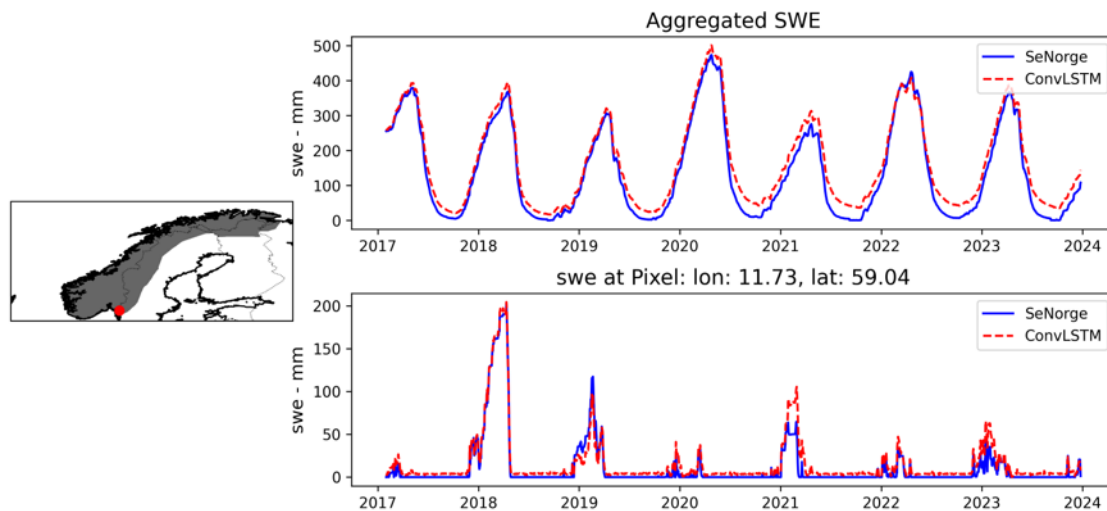


Figure 2.12. Temporal evaluation of the ConvLSTM model against the SeNorge snow model reference dataset over Norway for the period 2017–2024. The upper panel shows the domain-aggregated daily SWE time series, and the lower panel presents the SWE time series at a single representative pixel located in southeastern Norway.

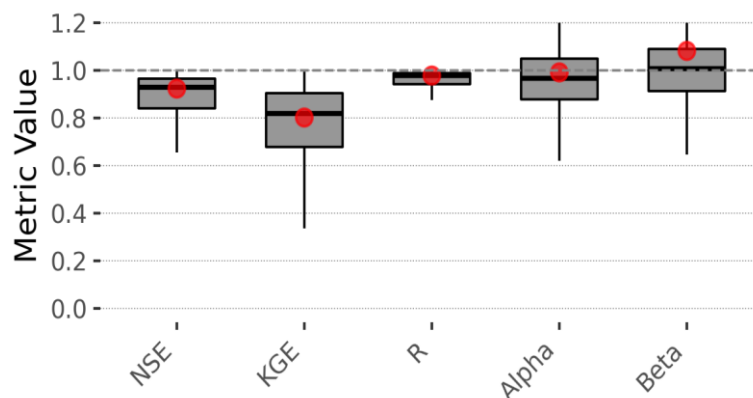


Figure 2.13. Summary of ConvLSTM model performance across Norway evaluated against the SeNorge snow model reference dataset. Red markers indicate median values. The dashed line at 1.0 marks ideal model performance. Red dots represent the mean values.

Figure 2.14 presents a spatial comparison of KGE performance metrics evaluated against in-situ observations for three different model configurations: SeNorge, ConvLSTM driven by ERA5 dataset (ConvLSTM-ERA5), and ConvLSTM driven by SeNorge dataset (ConvLSTM-SeNorge), at both daily and monthly temporal resolutions. At the daily scale, the SeNorge reference dataset achieves the strongest and most spatially consistent performance, with KGE values above 0.75 concentrated across southern and central Norway, as expected given that the in-situ observations are spatially closer to

the regions where SeNorge was originally calibrated. The ConvLSTM-SeNorge configuration closely mirrors the spatial performance pattern of the reference dataset, confirming that the surrogate model successfully emulates the SeNorge snow model even when validated against independent point-scale measurements. In contrast, the models also exhibit a broader spread of lower KGE values, particularly along coastal and northern stations, reflecting limitations in resolving local orographic and microclimatic effects. At the monthly temporal scale, performance improves notably across all three configurations, with a general shift toward higher KGE values and reduced spatial variability, suggesting that aggregating predictions to monthly timesteps effectively dampens the impact of short-term timing errors in snow accumulation and melt. Importantly, even at the daily scale, the ConvLSTM-ERA5 model retains acceptable skill at a substantial fraction of stations, demonstrating the transferability of the deep learning framework to globally available forcing data in the absence of high-resolution regional products.

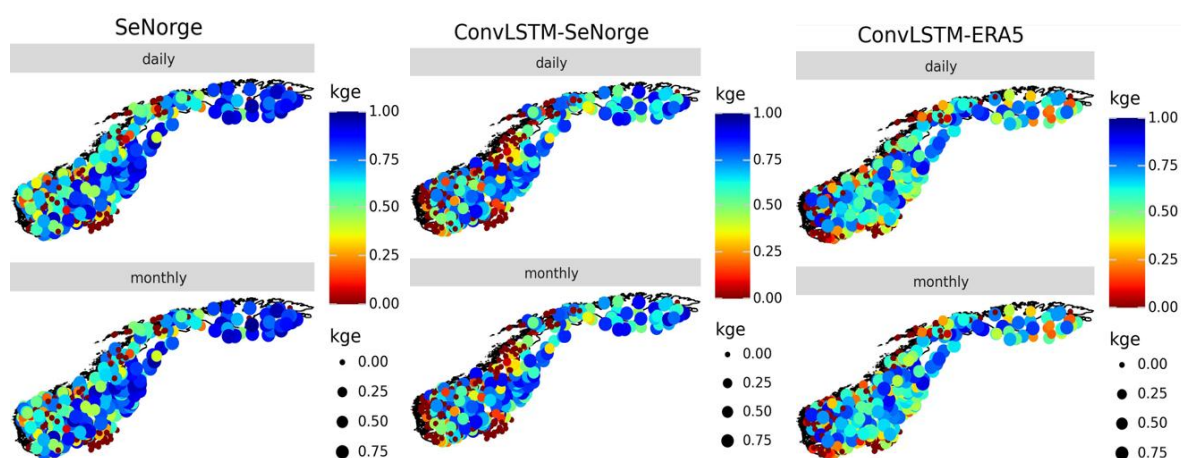


Figure 2.14. KGE comparison between simulated SWE derived from ConvLSTM and ERA5, and in-situ observations in Norway.

Figure 2.15 provides a quantitative synthesis of model performance against in-situ observations through boxplots of KGE and its three decomposed components for the SeNorge reference dataset and both ConvLSTM configurations. Note that we evaluated the simulated data for the whole period, including summer seasons. Regarding KGE, SeNorge achieves the highest and most consistent performance, with a median close to 0.5 and a relatively compact interquartile range, while ConvLSTM-SeNorge shows a similar median but with considerably wider spread, including outliers extending below -3, indicating occasional strong performance degradation at specific stations.

ConvLSTM-ERA5 yields a slightly lower median KGE but a more controlled distribution, suggesting that the ERA5-forced model, despite its coarser forcing, can produce performance slightly better in some regions compared to ConvLSTM-SeNorge. In terms of correlation, all three configurations perform comparably, with median values between 0.75 and 0.85, confirming that the temporal phasing of snow accumulation and melt is well captured regardless of the forcing dataset. However, the Variability Ratio reveals a systematic tendency toward overestimation of SWE variability in both ConvLSTM configurations, particularly ConvLSTM-SeNorge, where the interquartile range extends well above 1.0, suggesting that the model amplifies variability beyond what is observed at point scale. Similarly, the Bias Ratio indicates a positive systematic bias in both deep learning models relative to in-situ observations, contrasting with the more balanced behavior of the SeNorge reference. These discrepancies are partly attributable to the inherent scale mismatch between gridded model outputs and point-scale station measurements, where local snow redistribution processes driven by wind and topography are not fully resolvable at the model's spatial resolution.

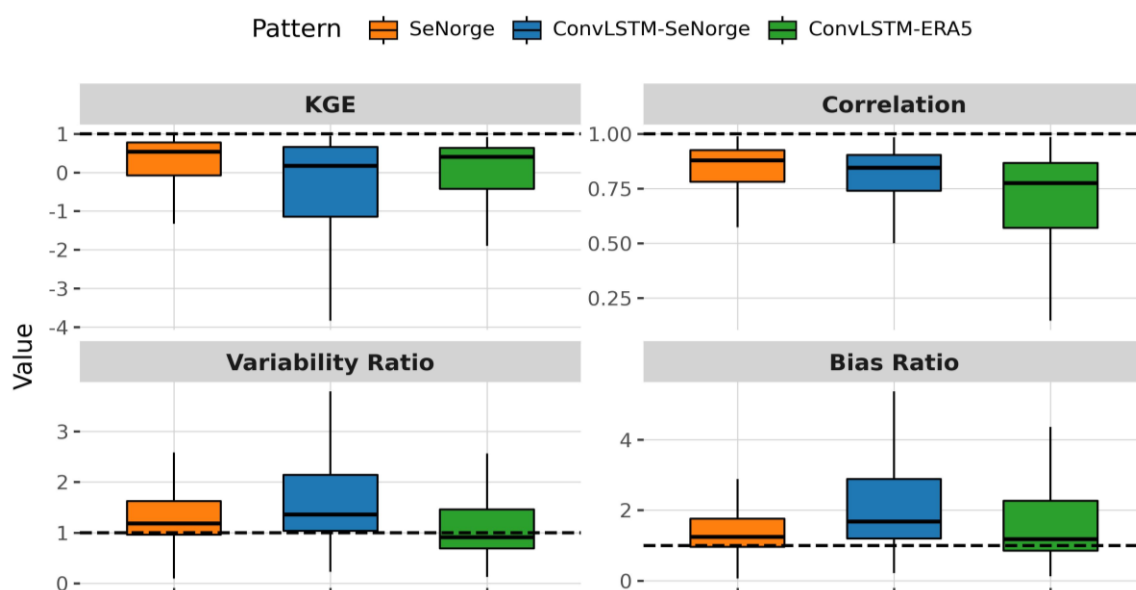


Figure 2.15. Summary performance of the comparison between simulated SWE derived from ConvLSTM and ERA5, and in-situ observations in Norway.

Alps

The ConvLSTM model trained over the Drammen river basin was subsequently applied to predict daily SWE across the European Alps. Unlike the transfer to the full Norwegian domain, we applied a fine-tuning stage prior to evaluation, given that the climatology and topographic conditions are slightly different compared to the observed in Norway. Fine-tuning was performed using simulated SWE fields from the LISFLOOD-EFAS reference dataset as the target variable, forced with EMO-1 meteorological inputs over the period 2010–2024. In this case, we adopted the period 2010–2015 for training, the period 2015–2018 for validation, and the period 2020–2024 for testing. Importantly, the same EMO-1 dataset is used as forcing for the LISFLOOD model itself, ensuring consistency between the reference simulation and the inputs provided to the ConvLSTM during the fine-tuning process. A key advantage of this approach is that fine-tuning does not require training the model from scratch. Instead, the pre-trained weights acquired during the original Norwegian training serve as the starting point, allowing the model to adapt its learned spatiotemporal representations to the specific climatic and orographic conditions of the Alpine domain with considerably fewer training epochs, reducing computational time compared to a full retraining procedure.

Figure 2.16 presents a spatially distributed evaluation of the ConvLSTM model performance across the Alps through four complementary metrics: bias percentage (Pbias %), Pearson Correlation, R^2 , and KGE, computed pixel-wise against the LISFLOOD-EFAS reference dataset. The Pbias map reveals a systematic overestimation tendency across most of the Alpine domain, with positive bias values predominantly ranging between 10% and 60%. This overestimation likely reflects the mismatch between the continental snow accumulation regime represented in the Norwegian training data and the distinct orographic precipitation patterns characteristic of the Alps, compounded by residual biases in the EMO-1 forcing that persist despite the quantile-quantile correction applied prior to model inference. The correlation map reveals moderate to strong temporal agreement across the Alpine domain, with values predominantly ranging between 0.6 and 1.0 across the higher-elevation interior regions, confirming that the model reliably reproduces the seasonal phasing of snow accumulation and melt cycles in the core Alpine areas. R^2 values follow a broadly similar spatial pattern, with

stronger performance concentrated in the higher-elevation areas, and a progressive reduction in skill toward the lower-elevation areas, where shallow and intermittent snowpacks pose a greater challenge for generalization. The KGE map also synthesizes these behaviors, showing values above 0.6 across portions of the high-Alpine core, with considerable spatial variability and localized deterioration across the southern and eastern margins of the domain. Collectively, these spatial performance patterns confirm that the ConvLSTM surrogate model, originally trained over the Drammen river basin retains meaningful predictive skill when transferred to the Alpine domain

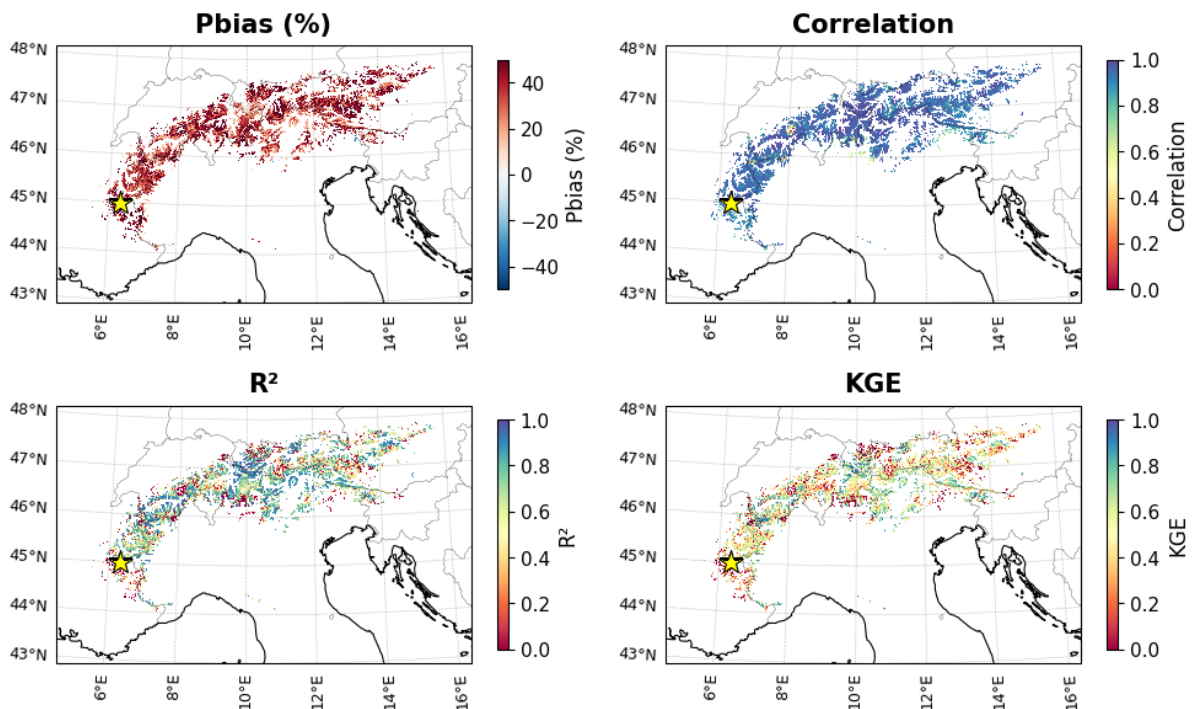


Figure 2.16. Spatial distribution of ConvLSTM model performance metrics evaluated against the LISFLOOD reference dataset across the Alps. Maps show Percentage Bias (Pbias, %), Pearson Correlation, coefficient of determination (R^2), and Kling-Gupta Efficiency (KGE), computed pixel-wise for the test period 2020–2024.

Figure 2.17 compares the domain-aggregated SWE and pixel-scale SWE at a representative Alpine location simulated over the 2020–2024 period. For both cases, the model reproduces the seasonal timing of accumulation and melt well correctly, tracking the onset of snowpack buildup each autumn and its recession each spring. However, a systematic positive bias is produced by ConvLSTM relative to LISFLOOD-EFAS, with the discrepancy most pronounced in 2024. This overestimation tendency is consistent with the positive Pbias identified in the spatial evaluation (Figure 2.16) and likely reflects the combined influence of residual bias inherited from the Norwegian training domain and local orographic effects not fully resolved by the EMO-1 forcing, an effect that intensifies during high-accumulation years and at finer spatial scales.

Figure 2.18 summarizes the overall performance of the ConvLSTM model across the Alps 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, with median NSE values approaching 0.7, indicating that the model explains a substantial fraction of the variance in SWE dynamics at the regional scale. Although a wider interquartile range and occasional outliers extending below zero reflect localized performance degradation in climatically challenging areas, such as low-elevation transitional zones and the southern region. KGE median values of approximately 0.5–0.6, combined

with a moderately widespread, confirm that the model maintains acceptable overall performance across the Alpine domain.

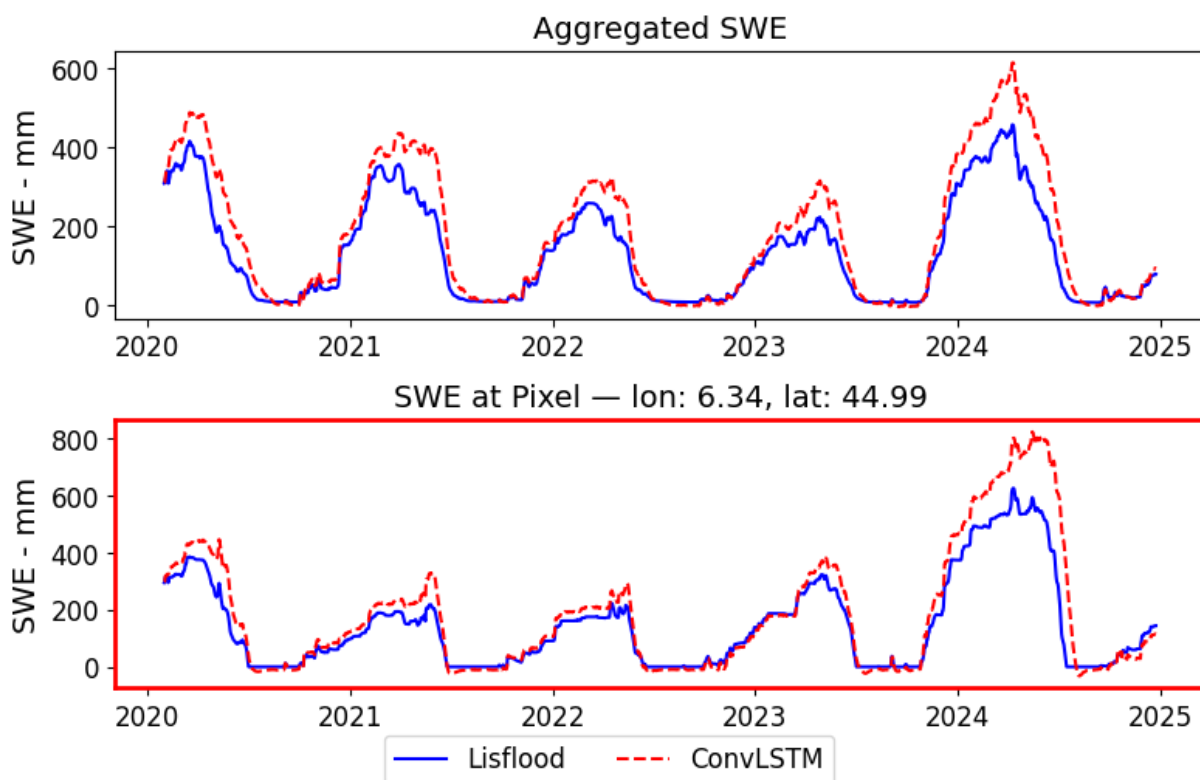


Figure 2.17. Temporal evaluation of the ConvLSTM model against the LISFLOOD reference dataset over Alps for the period 2020–2024. The upper panel shows the domain-aggregated daily SWE time series, and the lower panel presents the SWE time series at a single representative pixel located (lon: 11.73, lat: 59.04).

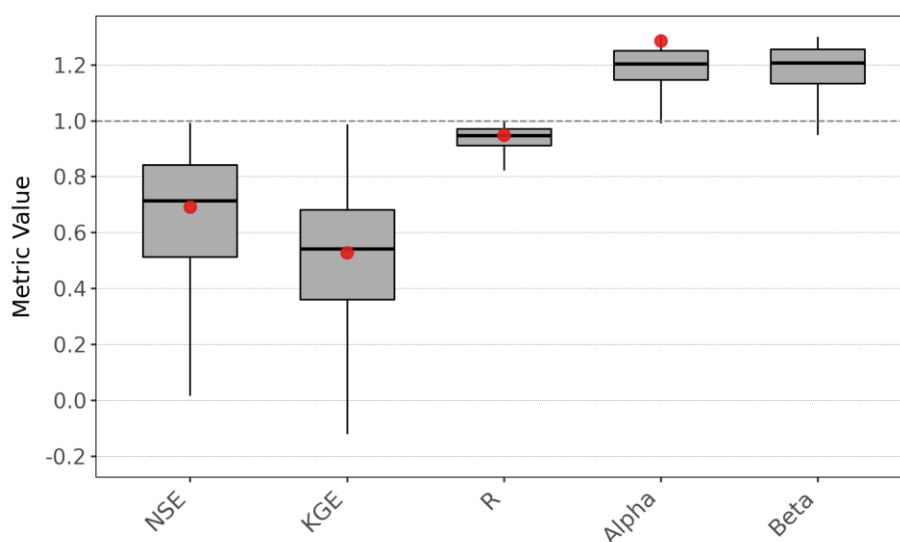


Figure 2.18. Summary performance of the comparison between simulated SWE derived from ConvLSTM and ERA5, and in-situ observations in the Alps. Red dots represent the mean value.

The correlation component (R) exhibits the tightest and most consistent distribution among all metrics, with median values close to 0.9 and minimal spread, reflecting the model's robust ability to reproduce the temporal phasing of snow accumulation and melt cycles across diverse Alpine landscapes. The α component shows a remaining tendency above 1.0, indicating that despite fine-tuning, a residual overestimation of SWE variability persists, likely associated with the coarser spatial representation of orographic precipitation gradients in the EMO-1 forcing relative to the complexity of Alpine terrain. Similarly, the β component clusters above 1.0, confirming a positive systematic bias in mean SWE magnitude that, while reduced by the fine-tuning procedure, has not been fully eliminated across the domain. Collectively, these results demonstrate that the combination of transfer learning and fine-tuning enables the ConvLSTM surrogate model to generalize effectively from the Drammen River Basin to the Alpine domain, preserving strong temporal predictive skill while substantially reducing, though not eliminating, the systematic biases in SWE variability and magnitude that arise from the domain shift.

Figure 2.19 and Figure 2.20 present a spatial comparison of KGE and Pearson correlation coefficients evaluated against in-situ observations from the NH-SWE dataset for both the ConvLSTM model and the LISFLOOD-EFAS reference simulation, at daily and monthly temporal resolutions across the Alpine domain. The KGE maps (Figure 2.19) reveal a markedly heterogeneous spatial pattern for both configurations, with values spanning the full range from poor to good performance scattered throughout the Alpine arc rather than following a clear elevation- or location-dependent gradient, indicating that station-specific factors play a larger role than broad spatial controls in determining KGE skill; this pattern is largely consistent between daily and monthly resolutions and between the two models. The Pearson correlation maps (Figure 2.20), by contrast, show consistently strong performance for both ConvLSTM and LISFLOOD-EFAS, with values above 0.75 dominating nearly the entire domain at both temporal resolutions, and only isolated stations, exhibiting weaker temporal agreement. This confirms that both models reliably capture the seasonal timing of snow accumulation and melt, regardless of forcing dataset or station location, with the spatial extent of high correlation values appearing broadly comparable between the surrogate and the reference simulation.

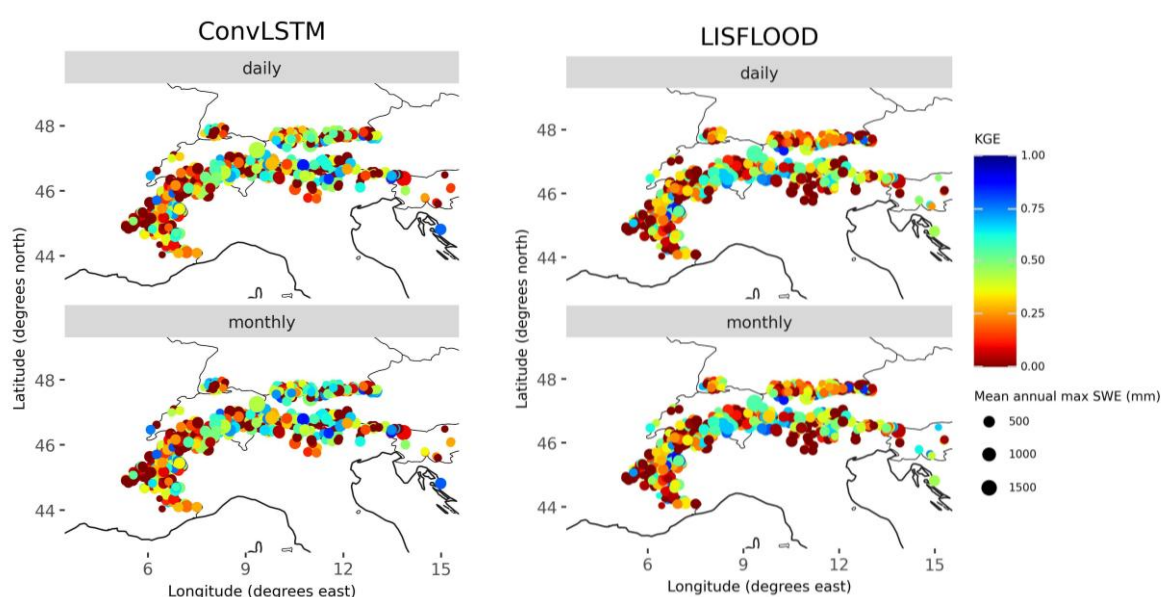


Figure 2.19. KGE comparison between simulated SWE derived from ConvLSTM and LISFLOOD, and in-situ observations in the Alps.

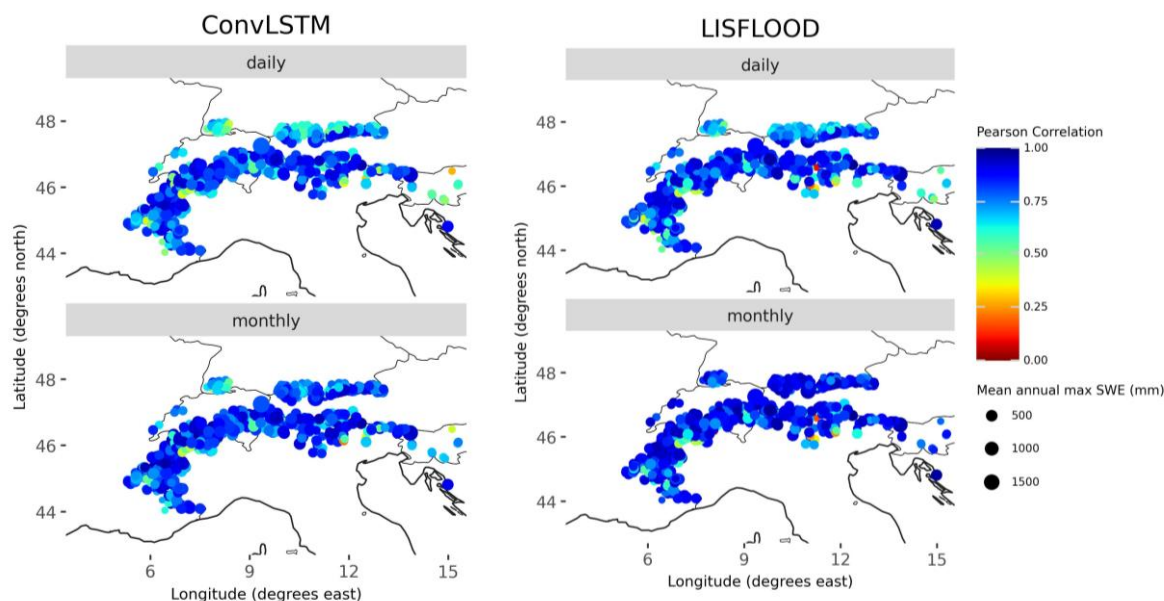


Figure 2.20. Pearson correlation comparison between simulated SWE derived from ConvLSTM and LISFLOOD, and in-situ observations in the Alps.

Figure 2.21 provides a quantitative synthesis of these results through boxplots of KGE and its three decomposed components, comparing LISFLOOD-EFAS and ConvLSTM at both daily and monthly resolutions. Median KGE is modest for both configurations (~ 0.35 – 0.40 daily, ~ 0.37 – 0.43 monthly), with ConvLSTM showing a slightly higher median but a wider interquartile range than LISFLOOD-EFAS, reflecting greater variability in performance across stations. Correlation remains the strongest and most consistent metric, improving from ~ 0.75 – 0.78 at the daily scale to ~ 0.80 – 0.87 at the monthly scale, with LISFLOOD-EFAS showing a marginally tighter distribution. Both the Variability Ratio and Bias Ratio cluster below 1.0 for both models, pointing to a systematic underestimation of SWE variability and magnitude relative to point-scale observations. Notably, ConvLSTM medians sit closer to the ideal value of 1.0 than LISFLOOD-EFAS, suggesting the surrogate represents amplitude and bias slightly more faithfully at the station scale. However, this aggregate improvement masks a systematic tendency for ConvLSTM to overestimate SWE relative to LISFLOOD-EFAS, indicating that the tighter median ratio reflects reduced spread rather than an absence of directional bias. Collectively, these results confirm that the ConvLSTM surrogate model performs comparably to, and in some respects exceeds, the LISFLOOD-EFAS reference across the evaluated metrics, demonstrating its potential as a computationally efficient alternative for basin-wide snow water equivalent assessment in the Alps.

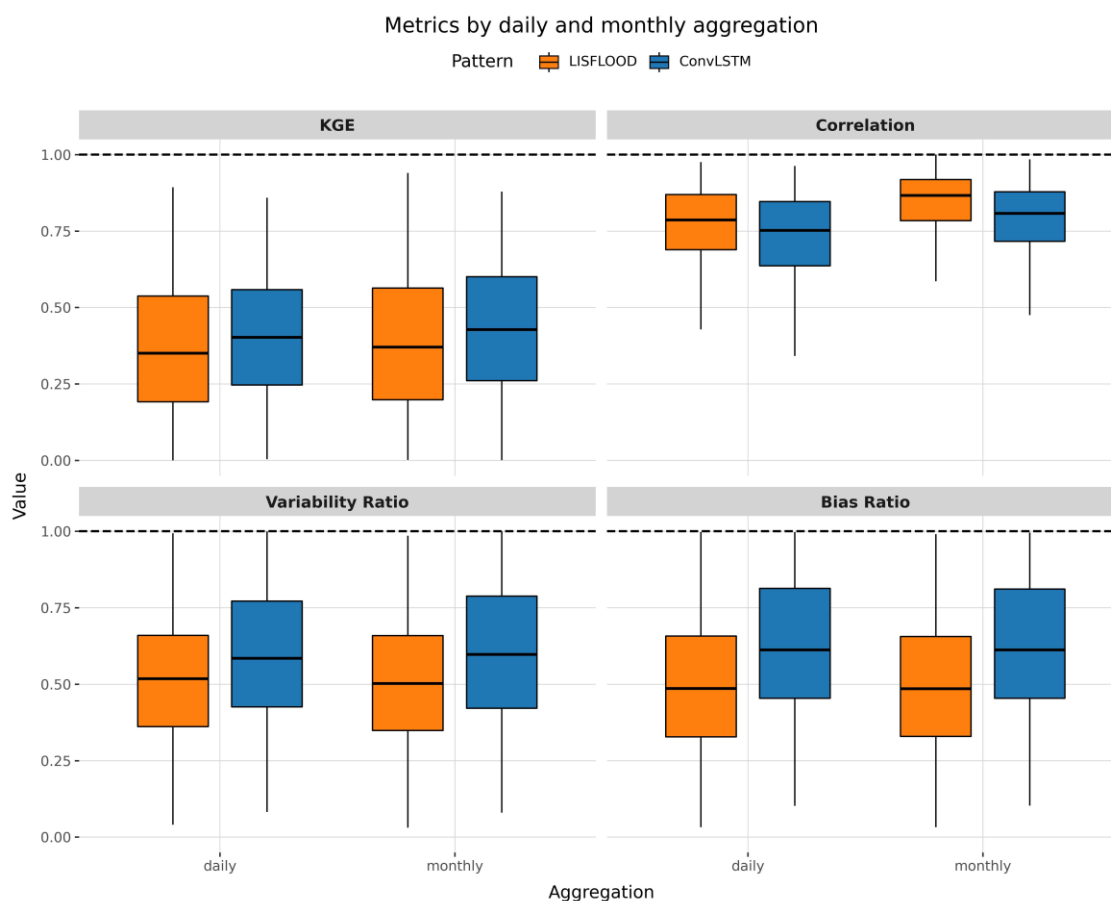


Figure 2.21. Summary performance comparison between simulated SWE derived from ConvLSTM and LISFLOOD, and in-situ observations in the Alps.

2.5. Conclusions

This chapter presented the development, evaluation, and spatial transferability of a ConvLSTM-based surrogate model for daily Snow Water Equivalent prediction across multiple European domains. The model was initially trained over the Drammen River Basin in Norway and subsequently transferred to the full Norwegian domain and the European Alps, demonstrating the potential of deep learning surrogates for large-scale snow hydrology applications. While the extension to the full Norwegian domain required no modification of the model architecture or weights, the transfer to the Alpine domain involved a targeted fine-tuning stage to address the climatological and orographic differences between the two regions.

Over the Drammen River Basin, the ConvLSTM model successfully reproduced the seasonal dynamics of SWE accumulation and melt, achieving KGE values above 0.75 when forced with the high-resolution SeNorge dataset, and above 0.6 when driven by the coarser ERA5-Land reanalysis. These results confirm that the model captures the primary spatiotemporal patterns of snowpack evolution while remaining robust to differences in forcing data resolution and quality. The implementation of a Bias-Aware Loss function and a scheduled sampling strategy during training contributed to reducing both systematic and random errors in the autoregressive prediction framework.

At the national scale over Norway, the ConvLSTM model generalized effectively without any modification of its architecture or weights, achieving median NSE values approaching 0.9 and KGE values of approximately 0.8 against the SeNorge reference dataset. The translational equivariance property of convolutional operations enabled seamless extension from the basin-scale training domain to the full national grid. Performance degradation was localized primarily in coastal and northern regions, where maritime climate influences and transient snowpack dynamics deviate substantially from the continental snow regime represented in the training data, identifying these as priority areas for targeted fine-tuning using local observations.

The transfer of the model to the Alpine domain represented a more challenging test of generalization, given the significant differences in orographic complexity, precipitation regimes, and snow accumulation dynamics relative to the Norwegian training environment. A fine-tuning stage was therefore applied using LISFLOOD-EFAS simulated SWE fields as the target variable, with EMO-1 meteorological inputs over the period 2010–2018. The same forcing dataset used by the LISFLOOD model itself, ensuring consistency between the reference and the ConvLSTM inputs. This fine-tuning procedure did not require any modification of the ConvLSTM architecture; only the model weights were updated through backpropagation on the Alpine training data, and the process converged in considerably fewer epochs than a full retraining from scratch. When evaluated against independent in-situ observations from the NH-SWE dataset, the ConvLSTM model performed comparably to the LISFLOOD-EFAS reference across KGE, correlation, variability, and bias metrics at both daily and monthly temporal resolutions, confirming its utility as a computationally efficient alternative for basin-wide SWE assessment.

These results demonstrate that the proposed ConvLSTM surrogate framework offers a flexible and transferable approach for large-scale snow water equivalent prediction in Europe, combining the computational efficiency of deep learning with the physical consistency of high-resolution regional reference datasets. Future work should focus on expanding the fine-tuning strategy to incorporate additional observation sources, exploring ensemble-based uncertainty quantification, and extending the framework to climate change projection scenarios to support long-term water resources planning across snow-dominated river basins.

3. Estimating daily water table depth

3.1. Introduction

Groundwater constitutes one of the largest accessible freshwater reservoirs on Earth and plays a fundamental role in sustaining ecosystems, agricultural production, industrial activities, and drinking water supplies. In many regions, groundwater also contributes significantly to river baseflow and helps maintain streamflow during dry periods, thereby supporting aquatic ecosystems and water availability during droughts. The state of groundwater resources is commonly characterized through Water Table Depth (WTD), defined as the vertical distance between the land surface and the groundwater table. WTD integrates the cumulative effects of precipitation, evapotranspiration, recharge processes, lateral groundwater flow, soil properties, and human water use, making it one of the most informative indicators of subsurface water availability and hydrological system resilience.

The importance of monitoring groundwater resources has increased considerably during the last decades due to the increasing water demand. Europe has experienced several severe drought episodes during recent years, including the prolonged events of 2018–2020 and 2022, which resulted in substantial reductions in soil moisture, groundwater storage, and river discharge (Boergens et al., 2020; Moravec et al., 2021). These events highlighted the vulnerability of groundwater systems to sustained precipitation deficits and increasing atmospheric evaporative demand, with drought impacts on crop yield reported to have tripled over the last decades in Europe (Brás et al., 2021). Moreover, consecutive droughts often prevent complete groundwater recovery between years, leading to cumulative subsurface water deficits that can persist for several years (Hartick et al., 2021). Consequently, accurate monitoring and forecasting of WTD have become essential components of climate-resilient water management strategies, supporting drought assessment, agricultural planning, groundwater protection, and adaptation to future climate conditions (Naumann et al., 2021).

Traditionally, groundwater dynamics have been studied using in-situ monitoring networks, remote sensing observations, and physically based hydrological models. While groundwater wells provide direct measurements of WTD, their spatial distribution is often sparse and heterogeneous, limiting their ability to represent regional groundwater variability. Satellite observations offer valuable information on large-scale terrestrial water storage variations, but they generally lack the spatial and temporal resolution required to resolve local groundwater dynamics (Dorigo et al., 2017). As a result, physically based integrated hydrological models have become one of the most widely used tools for estimating groundwater states and fluxes across large domains (Dasgupta et al., 2023).

Among these models, integrated hydrological models such as ParFlow/CLM have demonstrated remarkable capabilities for simulating the coupled interactions between the atmosphere, land surface, vadose zone, and groundwater system. These models solve the three-dimensional Richards equation for variably saturated subsurface flow while simultaneously representing land-atmosphere exchanges through physically based land surface schemes (Kollet and Maxwell, 2006; Kuffour et al., 2020). ParFlow applications have successfully reproduced groundwater dynamics across a broad range of spatial scales, from individual catchments to continental domains (Maxwell et al., 2015; Kollet et al., 2018; O'Neill et al., 2021). More recently, Belleflamme et al. (2023) developed a high-resolution ParFlow/CLM configuration (DE06) covering Germany and surrounding regions at approximately 0.6 km spatial resolution. Their results demonstrated that integrated hydrological modeling can accurately reproduce observed groundwater dynamics, with strong performance across several monitoring wells while simultaneously providing estimates of groundwater recharge, soil moisture, seepage water, capillary rise, and other hydrologically relevant variables.

An important barrier to using integrated hydrological models is that they remain computationally demanding. Large-scale simulations often require high-performance computing infrastructures, GPU acceleration, extensive atmospheric forcing datasets, and detailed descriptions of soil hydraulic properties, topography, and land cover (Hokkanen et al., 2021). Running these models at continental scales and high temporal resolutions can therefore become prohibitively expensive, particularly for operational forecasting systems or climate scenario analyses (Kollet et al., 2018). These challenges have motivated increasing interest in the development of surrogate modeling approaches capable of emulating the behavior of physically based models while substantially reducing computational costs.

This chapter presents a ConvLSTM-based surrogate modeling framework for estimating daily Water Table Depth across Europe. The model is trained on WTD simulations generated by the high-resolution ParFlow/CLM DE06 setup (Belleflamme et al., 2023) using precipitation and soil moisture information as primary predictors. Precipitation is represented through the Antecedent Precipitation Index (API), which summarizes the cumulative influence of previous rainfall events through an exponential decay formulation and provides a compact representation of groundwater recharge memory. Unlike the ConvLSTM framework presented in Chapter 2 for Snow Water Equivalent forecasting, the objective of this model is not to predict future groundwater conditions several days ahead, but rather to estimate current daily WTD states based on historical sequences of hydro-meteorological conditions. Since the original ParFlow/CLM simulations are available at very high spatial resolution, a patch-based training strategy was implemented to overcome memory limitations and efficiently learn local spatiotemporal groundwater dynamics. Subsequently, a transfer learning approach was applied to extend the trained framework from Central Europe to the entire European continent, using simulations from the Terrestrial Systems Modeling Platform (TSMP; Furusho-Percot et al., 2019) as an independent benchmark dataset. ERA5-Land reanalysis products were also employed as forcing variables to evaluate the transferability of the proposed framework using globally available meteorological datasets.

3.2. Study area and data

The study domain is presented in two stages. The initial training domain covers Central Europe, encompassing Germany and its neighboring countries, using WTD simulations from the high-resolution ParFlow/CLM DE06 setup (~0.6 km spatial resolution) as reference data (Figure 3.1). This region exhibits a marked altitudinal contrast between the northern and southern portions, ranging from the flat north European plain, where elevations near sea level are associated with shallow water tables and high groundwater-surface water connectivity, to the Alpine system in the southernmost sector, where elevations exceed 2500 m and steep hydraulic gradients promote rapid subsurface drainage, episodic recharge from snowmelt, and highly variable water table depths. A key advantage of using ParFlowCLM DE06 as the training reference is its combination of high spatial resolution and atmospheric forcing derived from the ECMWF HRES reanalysis product, which unlike fully coupled systems such as TSMP that internally compute all atmospheric variable, allows for a more reliable representation of the relationships between precipitation, soil moisture, and water table depth compared to observed data. Following the initial training, a transfer learning approach was applied to extend the ConvLSTM surrogate to the full European continent, using TSMP simulations over the CORDEX EUR-11 domain (0.11° resolution) as an independent benchmark dataset. The broader European domain introduces additional physiographic complexity, each characterized by distinct precipitation regimes, soil properties, and groundwater dynamics that deviate substantially from the Central European training environment. This strong continental-scale heterogeneity in topography and hydrogeological settings directly motivates the transfer learning strategy adopted in this study,

enabling the ConvLSTM surrogate to generalize its learned spatiotemporal representations of water table depth across diverse European landscapes with significantly reduced computational cost compared to full retraining from scratch.

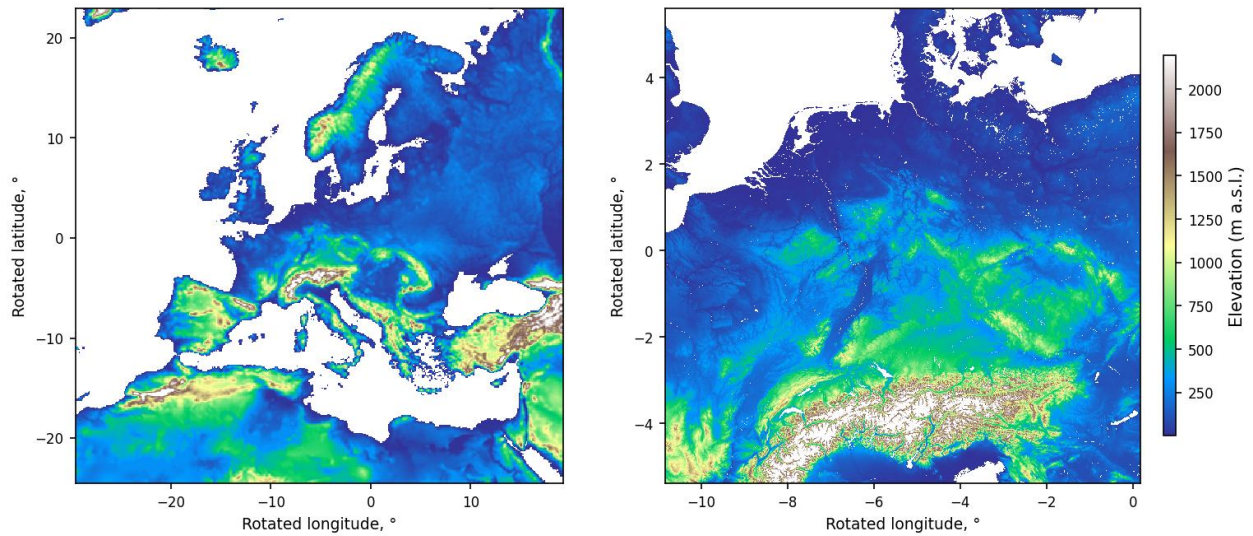


Figure 3.1. Topographic elevation for the two model domains used in this study. Left panel: pan-European TSMF domain on the CORDEX EUR-11 rotated grid (~12.5 km). Right panel: Central European ParFlow/CLM DE06 training.

This study used static input features to better represent the spatial heterogeneity of soil properties and groundwater conditions. Topography and slopes were derived from the ASTER Global Digital Elevation Model (GDEM), with river trajectories corrected using the hydrologically adjusted MERIT DEM to ensure realistic flow directions, particularly in flat terrain and channelized reaches. Soil texture information, including clay and sand content, was obtained from the SoilGrids250m global dataset (250 m resolution), which was used to classify soils into USDA texture classes.

Figure 3.2 and Figure 3.3 present the spatial distribution of these static input variables. These variables are of direct relevance to groundwater dynamics, as soil texture and terrain slope are primary controls on hydraulic conductivity, infiltration capacity, lateral subsurface flow, and ultimately water table depth.

Across both domains, clay content shows moderate to high values (15–35%) dominating central and western Europe, while the Iberian Peninsula and Mediterranean regions display higher clay fractions. Sand content exhibits a consistent and pronounced north-south gradient across both spatial scales, with values exceeding 70–80% across the northern lowlands of Scandinavia, the Netherlands, and northern Germany. The slope maps in both the east-west and north-south directions reveal near-zero values across the flat northern plains in both domains, where groundwater flow is predominantly vertical and driven by local recharge, while the Alpine foothills, the Pyrenees, the Carpathians, and the Scandinavian mountain ranges display strongly pronounced slope gradients, driving significant lateral groundwater redistribution, deepening water tables on hillslopes, and concentrating discharge in valley bottoms. While the coarser resolution of the CORDEX EUR-11 grid implies a smoother representation of terrain and soil variability compared to the DE06 domain, the broad spatial contrasts in hydraulic properties captured at both scales remain sufficient to guide the ConvLSTM surrogate in differentiating water table depth responses across the diverse hydrogeological settings present within each region, and their inclusion as static features in the ConvLSTM input is essential for enabling the model to generalize effectively from the Central European training domain to the full European continent.

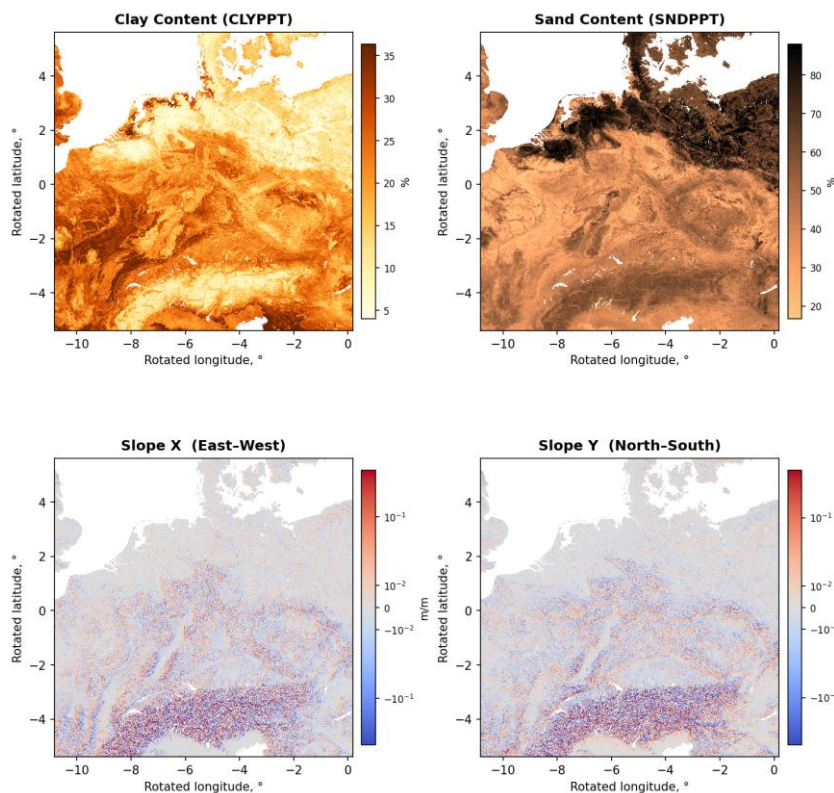


Figure 3.2. Spatial variability of clay content, sand content, slope x and slope y over Central Europe (ParflowDE06 domain).

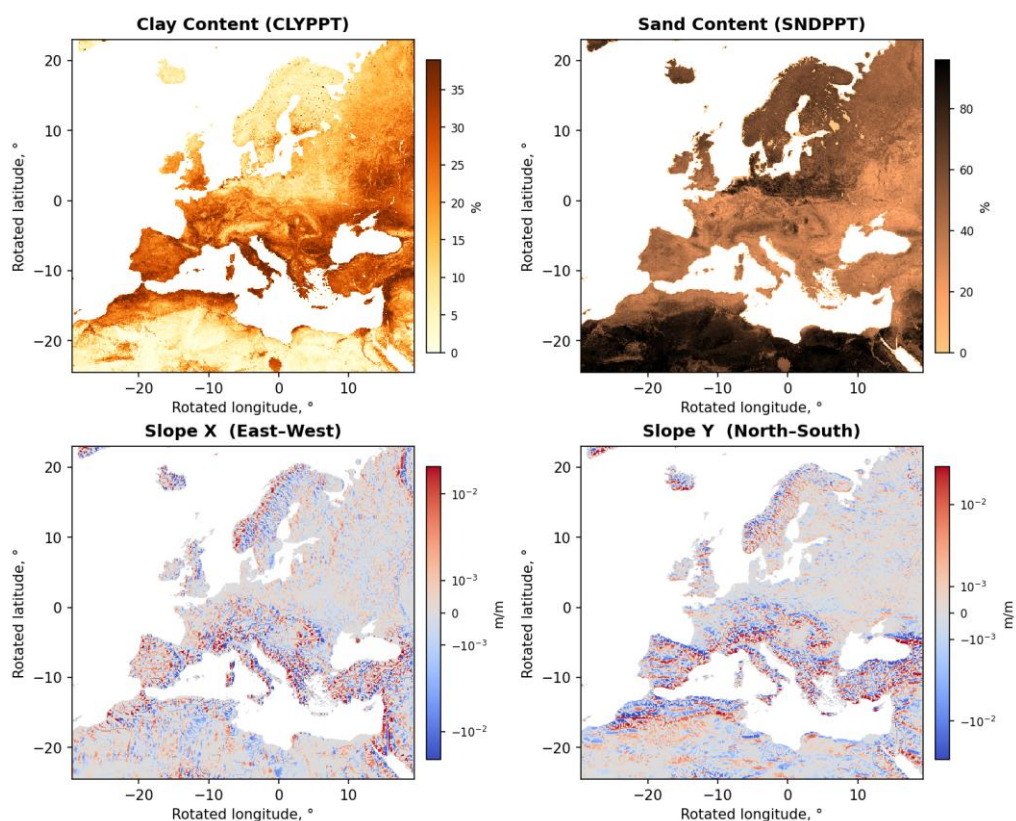


Figure 3.3. Spatial variability of clay content, sand content, slope x and slope y over Europe.

ParflowCLM

ParFlow is physics-based model that solves the Richards equation for variably saturated subsurface flow and applies the kinematic wave equation for overland flow generation (Kollet and Maxwell, 2006; Kuffour et al., 2020). It is coupled with the Common Land Model (CLM; Dai et al., 2003), which represents the exchanges of water, energy, and momentum between the land surface and the near-surface atmosphere, including processes such as evapotranspiration, snow accumulation and melt, and canopy interception. Together, ParFlowCLM simulates most of the surface and groundwater variables as a three-dimensional continuum, without requiring calibration, as the model parameterization is derived directly from physical properties of the soil and subsurface. An important advantage of this physics-based approach is that the governing equations remain valid across a wide range of hydroclimatic conditions, making the simulated WTD fields physically consistent and suitable as a training target for a data-driven surrogate model.

The DE06 setup operates on the DE-0055 grid, a rotated-pole equal-area Cartesian grid inscribed within the pan-European CORDEX EUR-11 domain. It covers Germany and its neighbouring regions with a horizontal resolution of $0.0055^\circ \times 0.0055^\circ$ (approximately $0.611 \text{ km} \times 0.611 \text{ km}$), resulting in a domain of $2,000 \times 2,000$ grid cells. The vertical dimension consists of 15 terrain-following layers with thickness increasing with depth, from 2 cm at the surface (uppermost layer) to 18 m for the lowermost layer, reaching a total soil depth of 60 m. This multi-layer structure allows DE06 to resolve the full soil column and the shallow saturated zone, providing vertically distributed information on soil moisture and groundwater dynamics.

Given that the ParFlow/CLM DE06 model setup is an operational model, it is forced with the deterministic medium-range weather forecast product HRES from the European Centre for Medium-Range Weather Forecasts (ECMWF; Owens and Hewson, 2018). The data, originally available at a coarser resolution of $0.1^\circ \times 0.1^\circ$ ($\sim 11 \text{ km}$), were spatially interpolated onto the DE-0055 grid ($\sim 0.611 \text{ km} \times 0.611 \text{ km}$) using the CDO (Climate Data Operators) bicubic remapping tool, and the corresponding hourly time steps were calculated by temporally interpolating the 3- or 6-hourly HRES output, since ParFlow/CLM requires atmospheric forcing at an hourly resolution. ParFlowCLM requires eight near-surface atmospheric variables as boundary conditions: 2-meter air temperature, specific humidity, zonal and meridional wind speed components at 10 m, surface air pressure, total precipitation, and visible and infrared downward radiation at the surface. All ECMWF forcing fields, originally available at approximately 11 km horizontal resolution, are interpolated onto the DE-0055 grid using bicubic remapping and temporally disaggregated to an hourly time step where needed.

TSMP

TSMP is a fully coupled terrestrial model closing the water and energy cycle from groundwater across the land surface to the top of the atmosphere (Gasper et al., 2014; Shrestha et al., 2014). TSMP comprises the hydrologic model Parflow for variably saturated groundwater and surface water flow, CLM for land surface processes and the numerical weather prediction model COSMO (Baldauf et al., 2011). The models are externally coupled via the Ocean Atmosphere Sea Ice Soil (OASIS3), allowing TSMP to run in fully coupled mode (Parflow-CLM-COSMO) or in hydrologic offline mode (Parflow-CLM). The TSMP model is set up over the European continent using a rotated latitude-longitude model grid with a horizontal resolution of 0.11° (12.5 km) from the CORDEX project. In the spin up, the groundwater-land surface subsystem was simulated using ParFlow-CLM using a 1979–1989 climatologic atmospheric forcing derived from the ERA-Interim reanalysis (Furusho et al, 2019). The ERA-Interim reanalysis data is also used for the COSMO model atmospheric initial and lateral boundary conditions for the EUR-11 domain throughout the fully coupled model simulations.

3.3. Methods

ConvLSTM framework

The proposed modeling framework for estimating daily WTD follows the same ConvLSTM-based architecture described in Chapter 2 for SWE prediction, with adaptations reflecting the distinct dynamics of groundwater systems (Table 3.1). Unlike the SWE model, which operates autoregressively over a 7-day forecast horizon, the WTD framework is designed to estimate the current daily groundwater state from a sequence of antecedent hydrometeorological conditions. The architecture consists of three main components: a static feature encoder, a dynamic feature encoder, and a ConvLSTM temporal integration module, followed by a convolutional decoder that produces the final WTD prediction map. Static inputs are encoded through a dedicated convolutional branch and replicated across all time steps, enabling the model to condition its predictions on the underlying physiographic and soil hydraulic heterogeneity of the domain. The temporal evolution of hydrological forcing is represented through two dynamic input variables at each time-step within the input sequence: the Antecedent Precipitation Index (API) and volumetric water content (VWC). Those data were also extracted from ParflowCLM simulations outputs and forcing data, respectively. The encoded static and dynamic features are concatenated along the channel dimension at each time step and passed through a multi-layer ConvLSTM module, which advances its hidden and cell states sequentially across the input window. Only the hidden state from the last time-step of the final ConvLSTM layer is retained and decoded into a single-channel WTD prediction map through successive convolutional layers with batch normalization and spatial dropout.

Table 3.1. Description of datasets and model configuration for the ConvLSTM framework used to estimated daily water table depth.

Component	Description
Static input channels	Topography, slope X, slope Y, clay content, sand content
Dynamic input channels	API, Volumetric water content
Input sequence length	30 days
ConvLSTM layers	2
Hidden dimension	64
Kernel size	3 x 3
Loss function	Hubber loss
Optimizer	Adam
Target variable	Water table depth (m) – ParflowCLM (DE05 domain)

Antecedent Precipitation Index (API)

The Antecedent Precipitation Index (API) is a scalar index that encodes the cumulative influence of prior precipitation events on soil moisture and groundwater recharge through an exponential decay function. The underlying premise is that the hydrological impact of past rainfall diminishes progressively with time, and that the rate of this decay is governed by soil drainage and evapotranspiration processes. API at time step t is computed as:

$API_t = k * API_{t-1} + P_t$	Equation 13
-------------------------------	-------------

where P_t is the precipitation on day t (mm) and k is a dimensionless decay coefficient ($0 < k < 1$) that controls the persistence of antecedent moisture conditions. Values of k close to 1 indicate slow drainage and long memory, characteristic of fine-textured or saturated soils, while lower values reflect rapid drainage typical of coarse-textured or sandy substrates. In this case, we used $k = 0.95$.

Su et al. (2021) demonstrated that the API is the single most important predictor of surface soil moisture in a global Random Forest framework, outperforming land surface temperature, vegetation indices, and soil texture variables. This finding underscores the dominant role of recent precipitation history in modulating near-surface hydrological states. In the context of the present ConvLSTM framework for Water Table Depth (WTD) estimation, API serves as the primary dynamic input variable, providing a compact and physically meaningful representation of the cumulative recharge signal driving groundwater fluctuations.

3.4. Results

Figure 3.4 presents the spatially distributed evaluation of the ConvLSTM model performance across the Central European training domain through four complementary metrics: KGE, Pearson Correlation, RMSE, and Percentage Bias (Pbias), computed pixel-wise against the ParFlowCLM DE06 reference dataset, while Figure 3.5 illustrates the temporal performance through domain-aggregated and pixel-scale WTD time series comparisons between ParFlowCLM and the ConvLSTM surrogate over the period 2012–2019.

The spatial performance maps reveal a heterogeneous but generally acceptable model behavior across the domain. KGE values above 0.6 dominate the flat northern lowlands of Germany and the Netherlands, where the strong coupling between antecedent precipitation and water table dynamics is well captured by the model, while progressive degradation toward red and orange tones is concentrated along river valleys, upland areas, and the Alpine foothills in the southern portion of the domain, where complex lateral groundwater flow and steep hydraulic gradients pose a greater challenge for the surrogate. The correlation map largely mirrors this spatial pattern, with values approaching 1.0 across the northern plains and deteriorating toward the more topographically complex central and southern regions, consistent with the known difficulty of convolutional architectures in resolving fine-scale subsurface flow processes driven by strong orographic gradients. The RMSE map also confirms this behavior, with errors below 0.5 m across the flat lowlands and increasing substantially in areas with deeper and more variable water tables associated with upland and Alpine terrain. The Pbias map reveals a mixed pattern of positive and negative biases largely within $\pm 10\%$, with a slight tendency toward underestimation in the southern uplands and overestimation in certain valley bottoms, suggesting that the model partially misrepresents the directionality of lateral groundwater redistribution in topographically complex areas.

Figure 3.5 illustrates the temporal performance of the ConvLSTM surrogate through domain-aggregated, and pixel-scale WTD time series and anomaly compared to the ParFlowCLM model over the period 2012–2019. At the domain-aggregated scale (upper panel), the ConvLSTM closely tracks both the seasonal cycle and interannual variability of ParFlowCLM WTD, accurately capturing the timing and magnitude of accumulation and recession cycles across all years. The anomaly bands (shaded areas, right axis) reveal that the surrogate reproduces the sign and approximate magnitude of departures from climatology throughout most of the record. At the individual pixel scale (lower panel), representative of a location with deeper and more dynamic water table behavior, the ConvLSTM reproduces the general seasonal pattern and tracks the anomaly sign correctly in most years, but exhibits occasional phase shifts and amplitude discrepancies, most notably an

underestimation of the positive anomaly peak around 2015–2016 and some divergence in the anomaly magnitude during the 2017–2018 drying period, suggesting that local groundwater dynamics at this location are not fully captured by the surrogate's antecedent precipitation and soil moisture inputs alone.

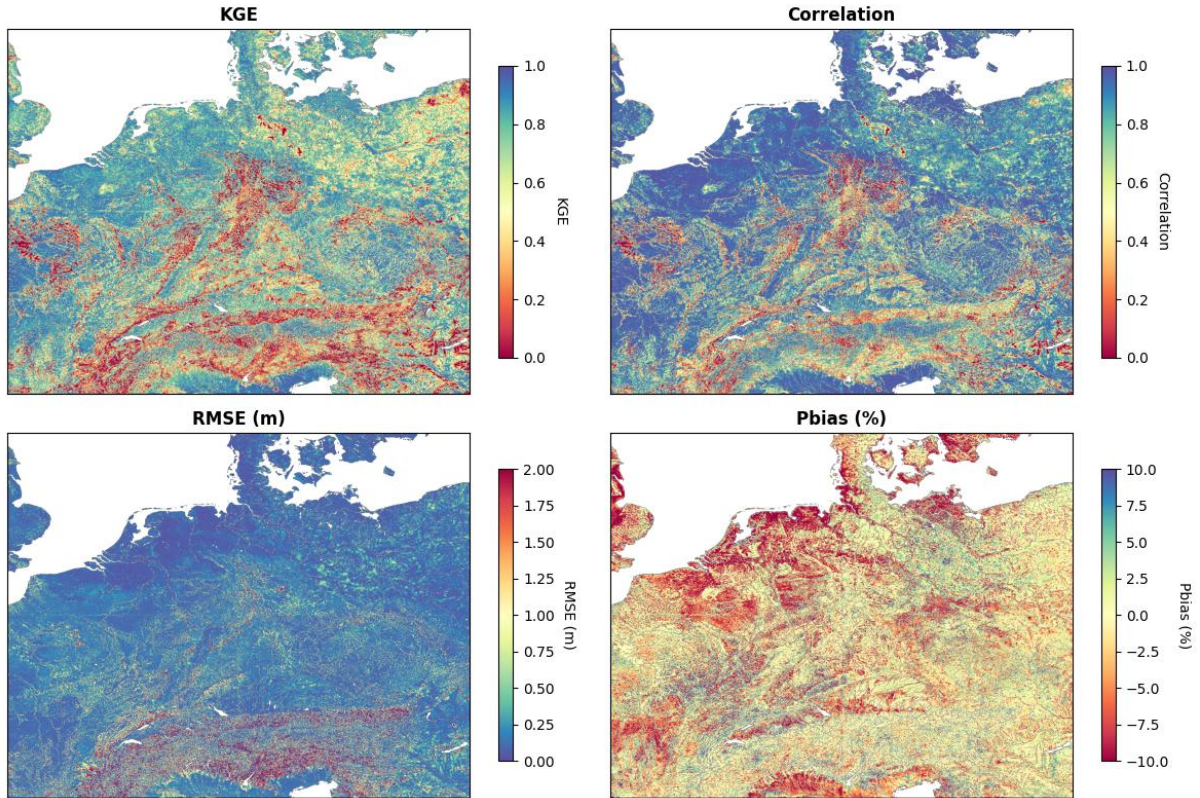


Figure 3.4. Spatial distribution of ConvLSTM model performance metrics evaluated against the ParFlow/CLM DE06 reference dataset across the Central European training domain.

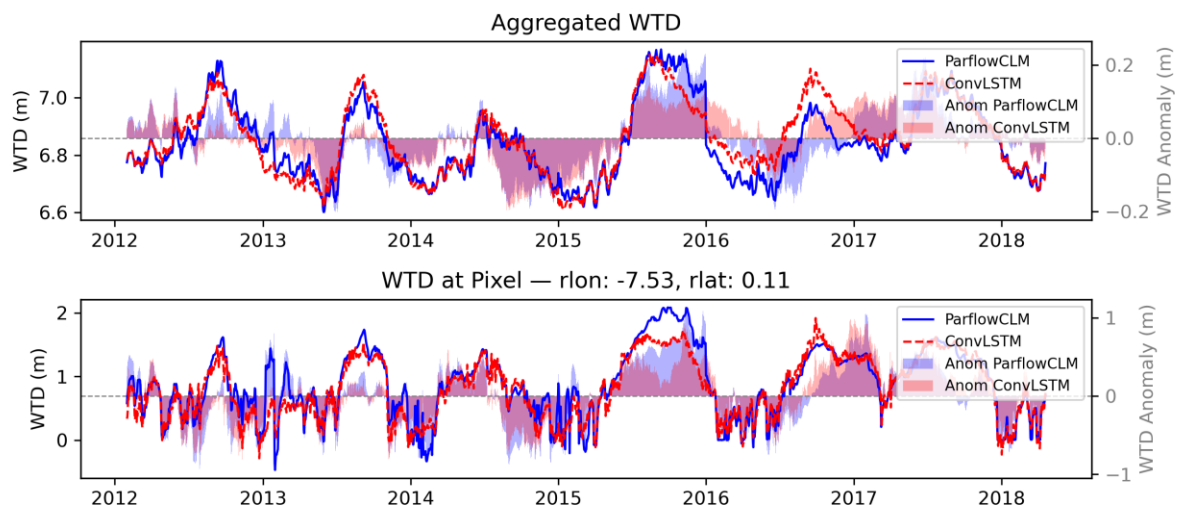


Figure 3.5. Temporal evaluation of the ConvLSTM model against the ParFlow/CLM DE06 reference dataset for the period 2012–2019. The upper panel shows the domain-aggregated daily WTD time series, and the lower panel presents the WTD time series at a single representative pixel.

Figure 3.6 summarizes ConvLSTM performance through boxplots of NSE, KGE, Pearson Correlation (r), variability ratio (a), and bias ratio (b) computed against the ParFlowCLM DE06 reference. The model achieves strong overall skill, with median NSE (~ 0.8), KGE (~ 0.75), and R (~ 0.9) confirming robust reproduction of WTD variance, temporal phasing, and overall hydrological behavior across the domain. The KGE components show a slight systematic underestimation of WTD variability amplitude (median a ~ 0.8), while b clusters tightly around 1.0, indicating minimal mean bias without any explicit bias correction. These results confirm that the primary remaining limitation of the surrogate lies in variability representation rather than in mean state or temporal dynamics.

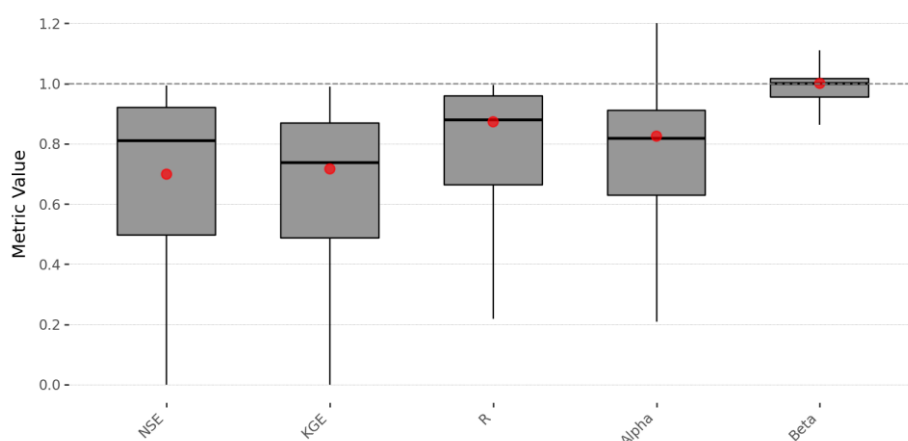


Figure 3.6. Summary performance of the comparison between simulated WTD derived from ConvLSTM and the ParFlowCLM model over the DE06 domain.

Figure 3.7 and Figure 3.8 present the spatially distributed performance metrics and temporal WTD time series and anomaly comparisons for the ConvLSTM surrogate when forced with ERA5-Land reanalysis data. Compared to the HRES-forced configuration, the ERA5-Land driven model shows a general degradation in spatial performance, with KGE values predominantly in the 0.2–0.6 range across most of the domain and a notably larger extent of red and orange tones, reflecting the limitations of the coarser ERA5-Land precipitation and temperature fields in resolving the fine-scale meteorological forcing that drives local groundwater recharge processes. The correlation map similarly shows a broader deterioration compared to the HRES-forced results, with values below 0.4 extending across a larger fraction of the topographically complex central and southern portions of the domain, suggesting that the temporal phasing of groundwater fluctuations is more strongly affected by forcing resolution than the mean state. The RMSE map reveals errors predominantly in the 0.5–1.5 m range across the northern lowlands, while the Pbias map exhibits a pronounced and spatially coherent positive bias dominating the northern and central portions of the domain, with values consistently reaching +7.5% to +10%, indicating a systematic overestimation of WTD that was not present in the HRES-forced configuration. These spatial patterns are clearly reflected in the temporal evaluation of Figure 3.8. At the domain-aggregated scale, the ConvLSTM forced with ERA5-Land reproduces the general seasonal cycle of ParFlow/CLM WTD but exhibits a persistent positive offset throughout the 2010–2020 period, consistent with the systematic Pbias identified in the spatial maps and clearly visible in the anomaly bands, where the ConvLSTM anomalies (red shading) remain systematically above those of ParFlow/CLM (blue shading) for most of the record, particularly from 2018 onward. At the individual pixel scale, the surrogate captures the seasonal cycle and the sign of interannual anomalies reasonably well, with the anomaly bands showing broad co-location of positive and negative departures from climatology in many years but exhibits amplified variability and frequent

overestimation of peak WTD values during wet seasons, producing anomaly excursions that exceed those of the reference model. These results demonstrate that while the ConvLSTM surrogate retains meaningful skill in reproducing the phasing of groundwater fluctuations when forced with globally available ERA5-Land reanalysis data, the systematic positive bias and amplified variability introduced by the coarser forcing represent the primary limitations of this configuration, identifying bias correction of ERA5-Land precipitation as a priority preprocessing step for improving the transferability of the framework to domains where high-resolution regional forcing products are unavailable.

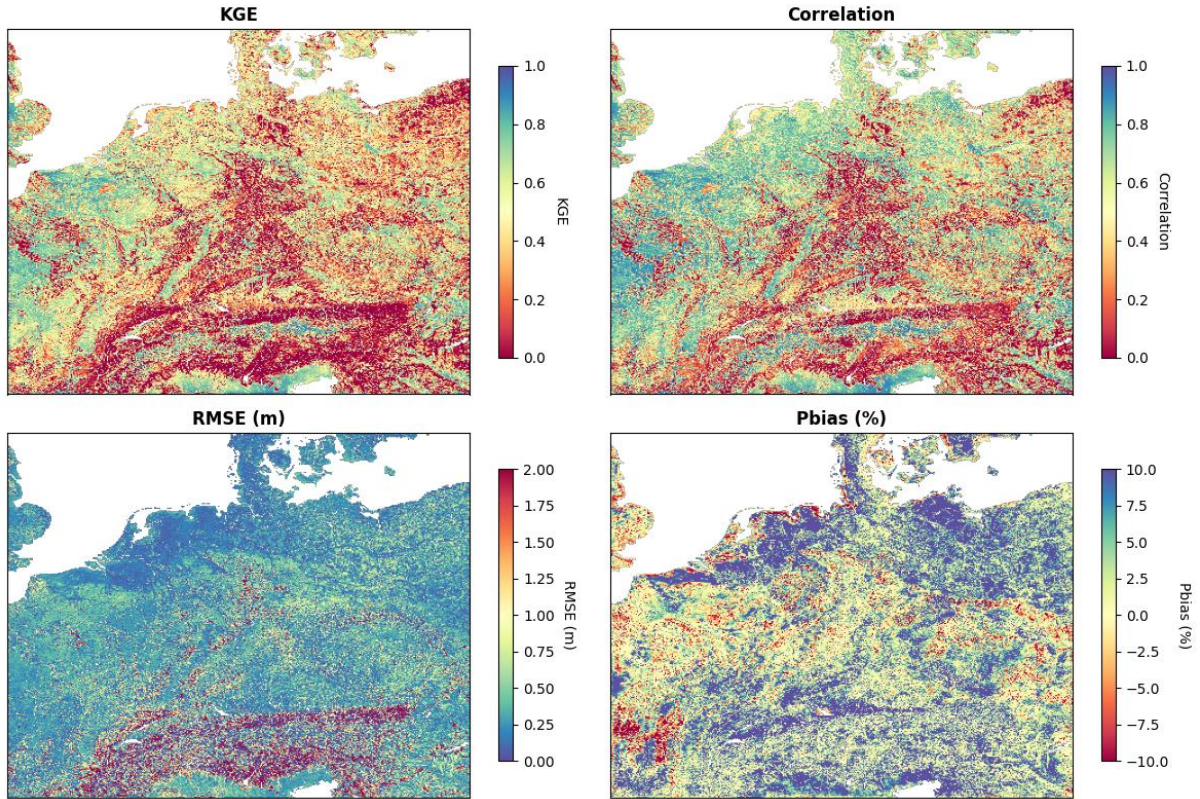


Figure 3.7. Spatial distribution of ConvLSTM model performance metrics evaluated against the ParFlow/CLM DE06 reference dataset across the Central European training domain.

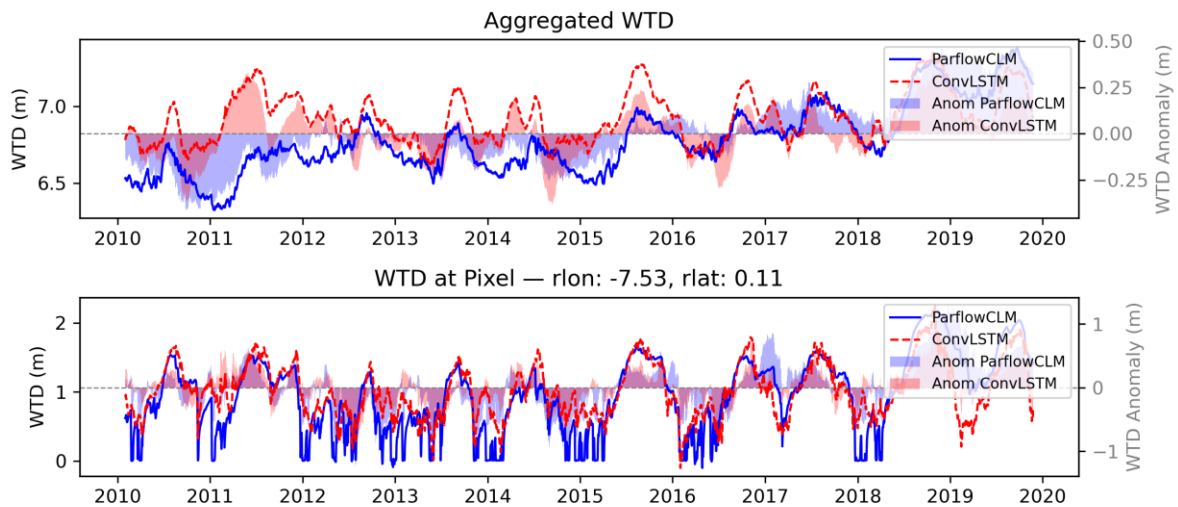


Figure 3.8. Temporal evaluation of the ConvLSTM model against the ParFlow/CLM DE06 reference dataset for the period 2010–2020. The upper panel shows the domain-aggregated daily WTD time series, and the lower panel presents the WTD time series at a single representative pixel located.

Figure 3.9 summarizes the ConvLSTM performance when forced with ERA5-Land reanalysis data through boxplots of NSE, KGE, Pearson Correlation (r), variability ratio (α), and bias ratio (β), computed against the ParFlow/CLM DE06 reference. Compared to the HRES-forced configuration, all metrics show a clear and systematic degradation, with median NSE dropping to ~ 0.45 , KGE to ~ 0.5 , and R to ~ 0.6 , confirming that forcing data resolution represents the primary limiting factor for surrogate model performance. The component (~ 0.8) reveals a systematic underestimation of WTD variability amplitude, while remains tightly clustered around 1.0, indicating that despite the overall degradation, the model preserves the mean WTD magnitude with minimal bias. These results confirm that ERA5-Land forcing errors manifest primarily as temporal and variability distortions rather than as a shift in the mean groundwater state, underscoring the importance of high-resolution meteorological forcing for accurately reproducing WTD spatiotemporal dynamics across Central Europe.

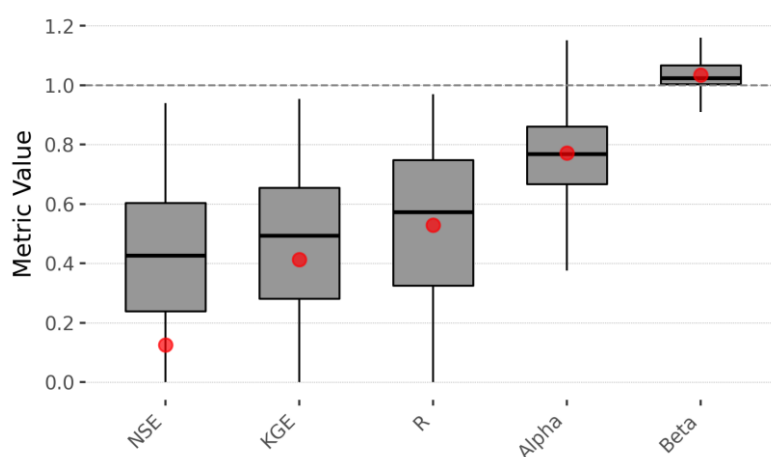


Figure 3.9. Performance summary of the ConvLSTM surrogate forced with ERA5-Land and evaluated against the ParFlow/CLM DE06 reference over the Central European domain. Boxplots show the distribution of NSE, KGE, Pearson correlation, variability ratio, and bias ratio across valid grid cells; the dashed reference line indicates ideal model performance where applicable. Red dots represent the mean value across the distribution.

Figure 3.10 shows Pearson correlation coefficients against in-situ WTD observations for ParFlow/CLM and ConvLSTM, with time series at six selected stations (2011–2016). Both models successfully capture WTD anomalies across different regions of Central Europe, as reflected in the broadly positive correlations distributed across the domain. However, ParFlow/CLM achieves systematically stronger correlations with observed values than the ConvLSTM surrogate, without a clear geographical distinction between regions of better and worse performance in either model. Negative correlations at isolated stations in both configurations, likely reflect anthropogenic influences or complex local hydrogeological conditions not captured by either uncalibrated model.

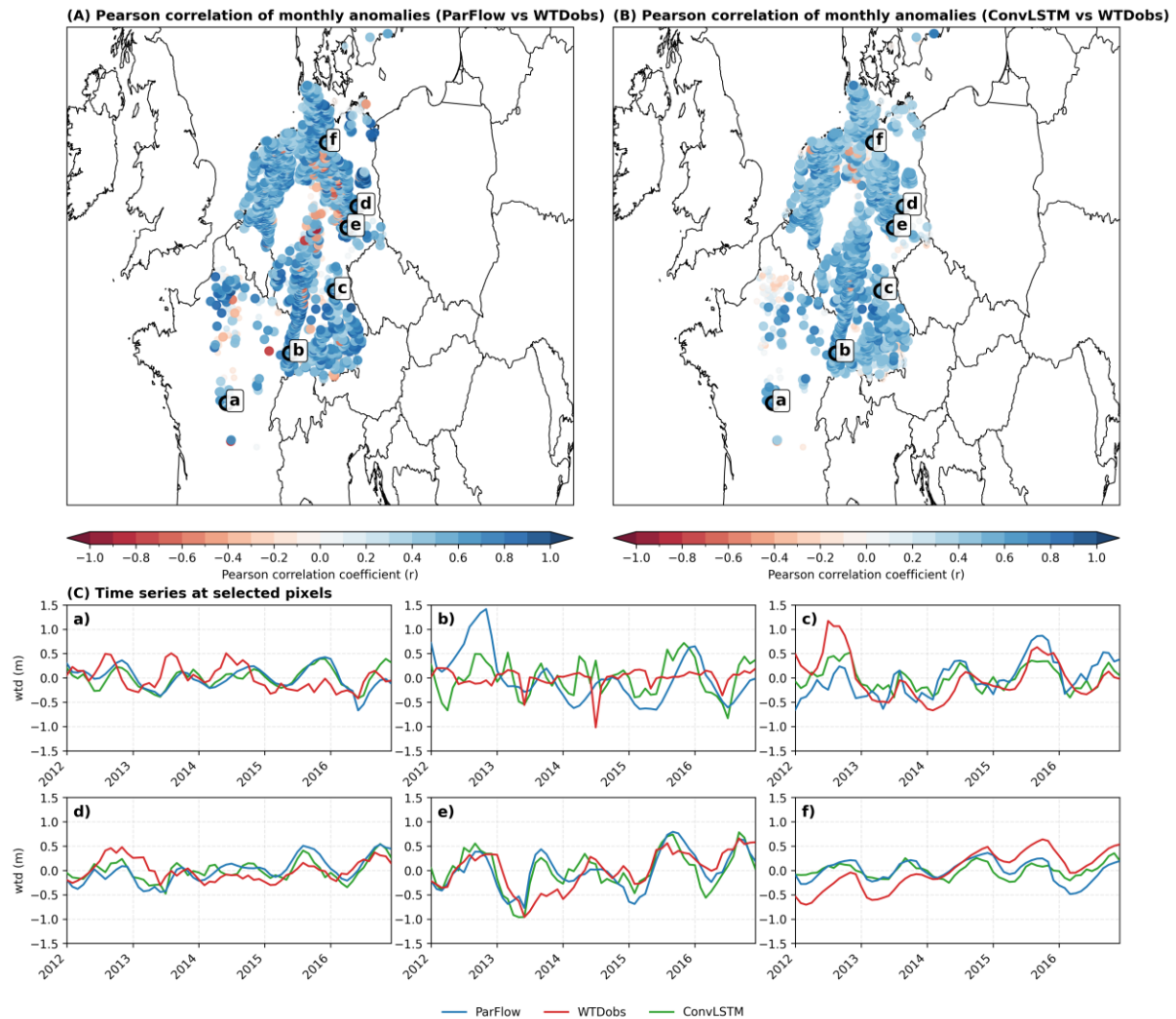


Figure 3.10. Pearson correlation of monthly WTD anomalies between simulated and in-situ observed water table depth in Central Europe. (A) Spatial distribution of the Pearson correlation coefficient (r) between ParFlow/CLM DE06 and in-situ observations (WTDobs). (B) Spatial distribution of the Pearson correlation coefficient (r) between the ConvLSTM surrogate and WTDobs at the same well locations. (C) Time series of monthly WTD anomalies at six representative stations over the period 2012–2016.

Europe

The ConvLSTM was subsequently transferred to the full pan-European domain using TSMP simulations over the CORDEX EUR-11 grid as an independent benchmark dataset, without applying any fine-tuning of the model weights (Figure 3.11). The spatial evaluation of this transfer reveals a clear and physically interpretable performance gradient driven by the underlying continental-scale variability in subsurface architecture and hydroclimatic forcing.

The strongest model skill, with KGE values approaching 0.8–1.0, Pearson correlation near unity, and RMSE below 0.4 m, is concentrated across the North European Plain, encompassing northern Germany, the Netherlands and Scandinavia, where shallow, precipitation-responsive water tables in highly permeable glacial and fluvioglacial sandy deposits ensure that the Antecedent Precipitation Index provides a near-complete representation of the recharge signal driving groundwater fluctuations. Performance degrades systematically across the major mountain massifs of Europe, including the Alps, the Carpathians, and the Balkan ranges, where elevated RMSE values reflect both the structural absence of snowmelt as an explicit forcing variable and the increasing climatic distance

from the training domain, causing the surrogate to misrepresent the timing and magnitude of delayed recharge pulses that dominate groundwater dynamics in snow-influenced catchments. The Pbias map further reveals a clear latitudinal gradient in the direction of systematic error: a pronounced positive bias (overestimation), reaching values of +6% to +10%, dominates Scandinavia and the northernmost parts of the domain, whereas a distinct and diagnostically informative pattern emerges across the Iberian Peninsula and Mediterranean coastal zones, where relatively high temporal correlation coexists with a pronounced negative Pbias (−6% to −10%), indicating that while the model successfully reproduces the phasing of seasonal groundwater fluctuations, it systematically underestimates WTD magnitudes there, likely reflecting the mismatch between the deep, geologically controlled aquifer systems characteristic of these semi-arid regions and the shallower, recharge-dominated groundwater dynamics represented in the Central European training domain, and highlighting these regions as priority candidates for a targeted fine-tuning stage using local reference simulations or observations.

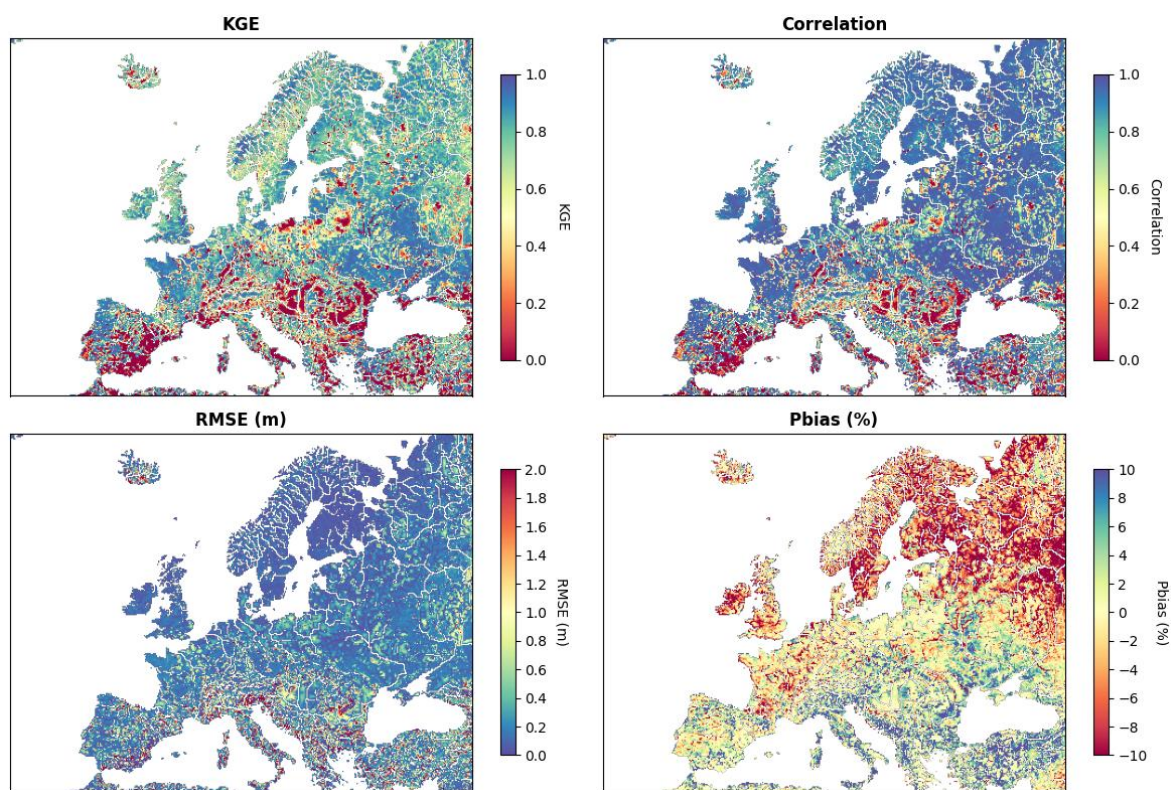


Figure 3.11. Spatial distribution of ConvLSTM model performance metrics evaluated against the TSMP reference dataset across the pan-European domain. Maps show KGE, Pearson correlation, RMSE, and percentage bias, highlighting regional differences in transferability from the Central European training domain to the full European domain.

The temporal evaluation and summary performance metrics presented in Figure 3.12 and Figure 3.13, respectively. The results reveal that the direct transfer of the ConvLSTM to the full pan-European TSMP domain achieves a level of predictive skill that is both encouraging and physically meaningful, particularly considering that no fine-tuning of model weights was performed prior to inference. At the domain-aggregated scale, the ConvLSTM reproduces the seasonal cycle of water table depth, accurately tracking the timing of annual recession and recovery phases, suggesting that the surrogate inherits a mild but systematic overestimation tendency that likely originates from the mismatch between the precipitation climatology of the Central European training environment and the broader pan-European domain. Importantly, the model captures the interannual variability of aggregated WTD

with high accuracy, correctly reproducing the anomalously deep-water tables of the 2018 drought year without any domain-specific calibration, which constitutes a particularly stringent and consequential test of transferability given the non-stationarity of hydroclimatic extremes. At the individual pixel scale, representative of a location in the western European domain, the surrogate tracks the high-frequency seasonal fluctuations and the progressive multi-year drying trend with reasonable accuracy, though with slightly amplified variability relative to the TSMP reference, consistent with the Alpha component in the boxplot summary showing a median variability ratio near 0.85, indicating a modest but systematic underestimation of WTD amplitude at the local scale.

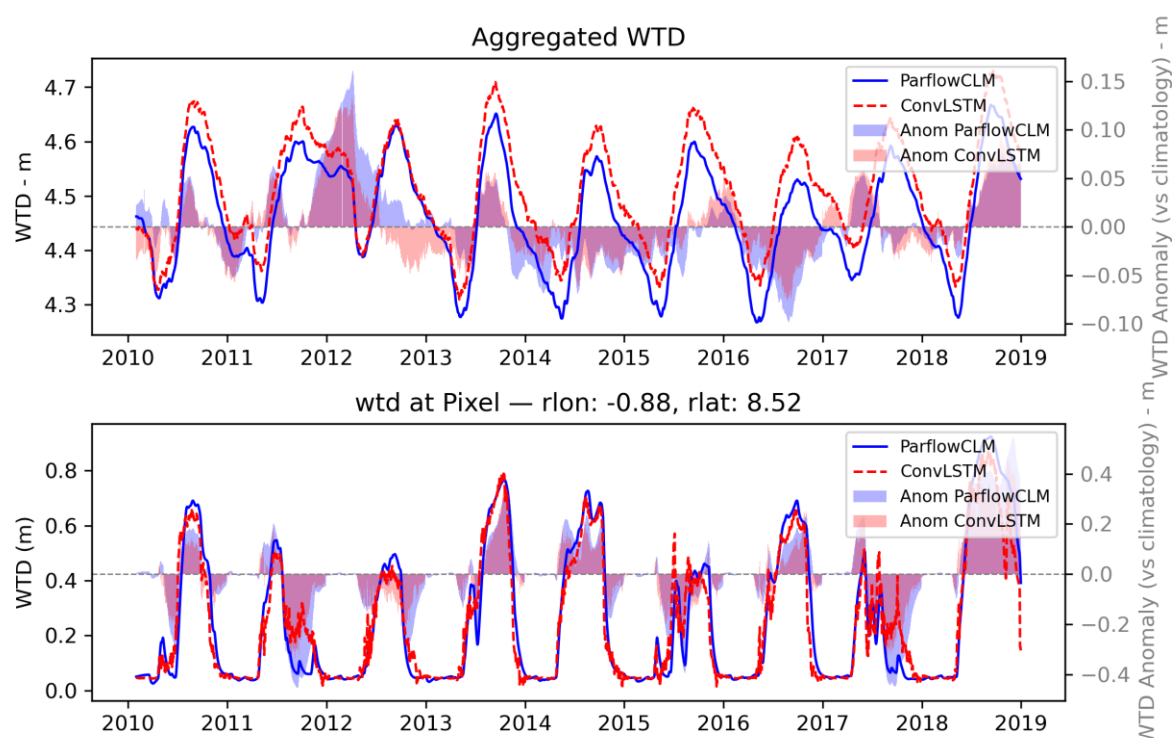


Figure 3.12. Temporal evaluation of the ConvLSTM model against the TSMP domain for the period 2010–2019. The upper panel shows the domain-aggregated daily WTD time series, and the lower panel presents the WTD time series at a single representative pixel located.

The summary boxplots further consolidate these findings: median NSE approaching 0.8 and KGE near 0.8 confirm that the surrogate model explains the vast majority of WTD variance across the European domain without retraining, while the Pearson correlation median exceeding 0.9 underscores the robustness of temporal phasing reproduction across diverse hydrogeological settings. The Beta component clustering tightly around 1.0 with minimal spread is particularly noteworthy, as it demonstrates that despite the domain shift from Central Europe to the full continental scale, the ConvLSTM preserves the mean WTD magnitude across the domain without introducing a systematic mean bias, suggesting that the static soil property inputs, successfully anchor the surrogate's predictions to the correct mean groundwater state even in regions not represented during training. Collectively, these results establish that the proposed ConvLSTM framework possesses genuine and transferable hydrological knowledge that generalizes beyond its training boundaries, while also identifying variability representation and localized positive offsets as the primary targets for improvement through a targeted fine-tuning strategy.

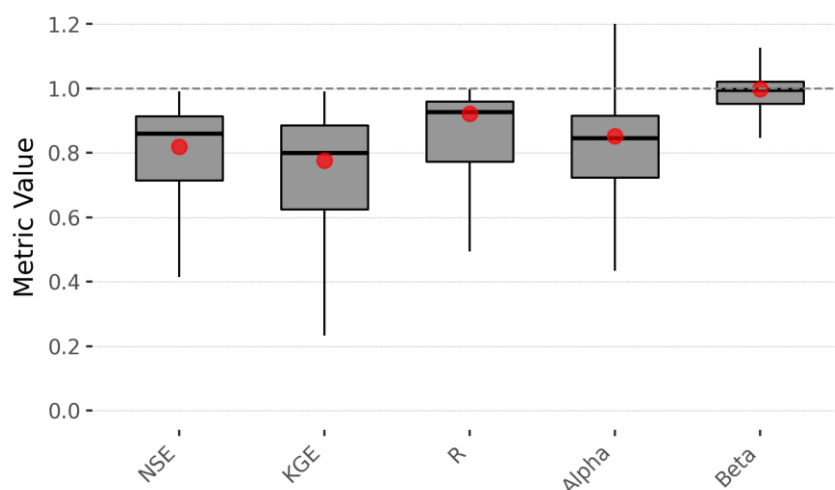


Figure 3.13. Summary performance of the comparison between simulated WTD derived from ConvLSTM over Europe.

When the ConvLSTM surrogate is forced with ERA5-Land reanalysis data instead of the original TSMP meteorological inputs, and without applying any prior bias correction to the forcing fields, a systematic and spatially coherent degradation in predictive skill is observed across the pan-European domain, confirming that forcing data quality represents the primary limiting factor for surrogate model performance at the continental scale (Figure 3.14). Spatially, KGE values drop substantially across central and western Europe, with the most severe deterioration concentrated over the British Isles, France, and the Mediterranean coastal regions, where the coarser spatial structure of ERA5-Land precipitation fields fails to resolve the fine-scale atmospheric forcing patterns governing local groundwater recharge, while the northern European lowlands of Scandinavia and the eastern European plain retain relatively acceptable performance with RMSE values below 0.6 m, consistent with the dominance of vertically driven, precipitation-responsive water table dynamics in these sandy environments that remain partially resolvable even at ERA5-Land resolution.

Temporally, while the ConvLSTM reproduces the general seasonal cycle of water table depth throughout the 2010–2019 period, it exhibits notable phase shifts and amplitude distortions that were largely absent in the TSMP-forced configuration, with the surrogate frequently anticipating or delaying seasonal recession and recovery peaks by several weeks, and the 2018 drought signal, appearing attenuated and temporally displaced, underscoring the sensitivity of extreme event reproduction to forcing data quality (Figure 3.15).

The summary performance boxplots (Figure 3.16) consolidate these observations quantitatively, with median NSE dropping to approximately 0.45, KGE to around 0.4, and Pearson correlation to nearly 0.5, representing a substantial reduction compared to the TSMP-forced medians, while the Beta component remaining tightly clustered around 1.0 demonstrates that despite the overall degradation the surrogate preserves the correct mean WTD magnitude, and the Alpha median near 0.8 reveals a persistent underestimation of variability amplitude that collectively identifies the application of quantile-quantile bias correction to ERA5-Land meteorological fields as the single most impactful preprocessing step for recovering predictive skill in the operational deployment of the framework across climatically diverse European domains.

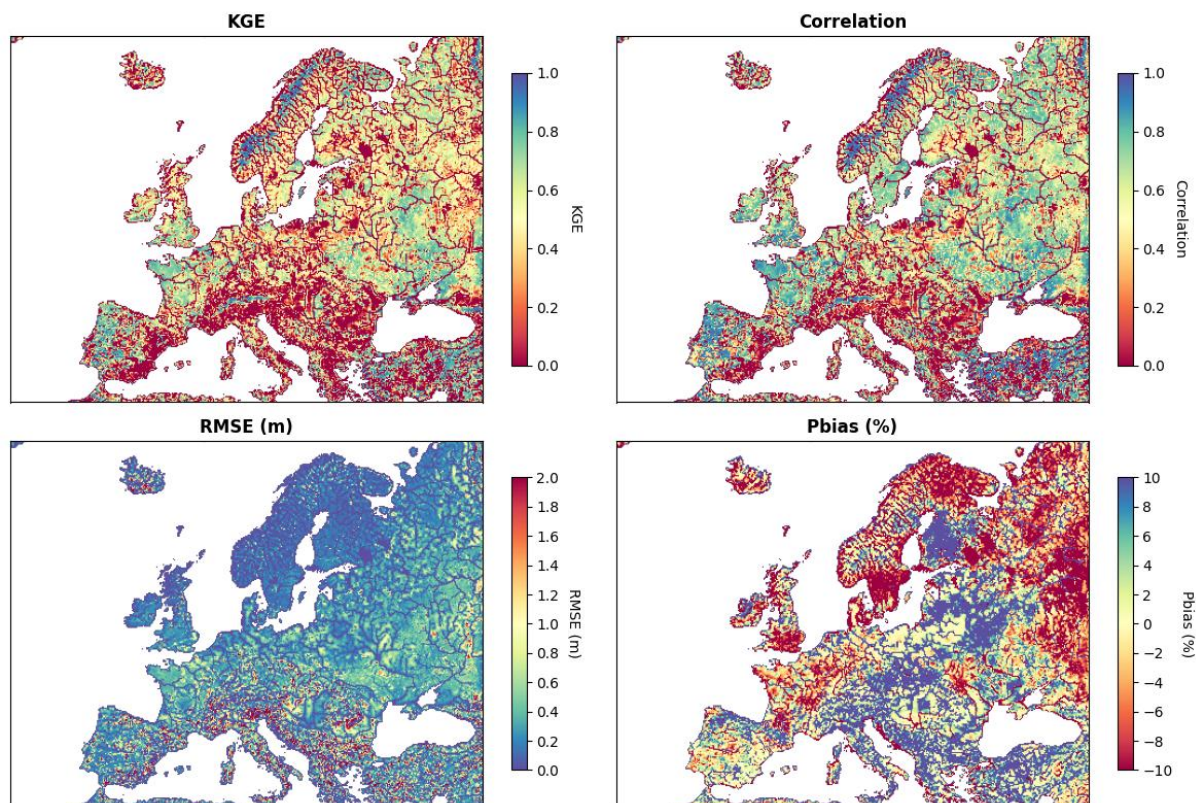


Figure 3.14. Spatial distribution of ConvLSTM model performance metrics, driven by ERA5 data, evaluated against the TSMP domain.

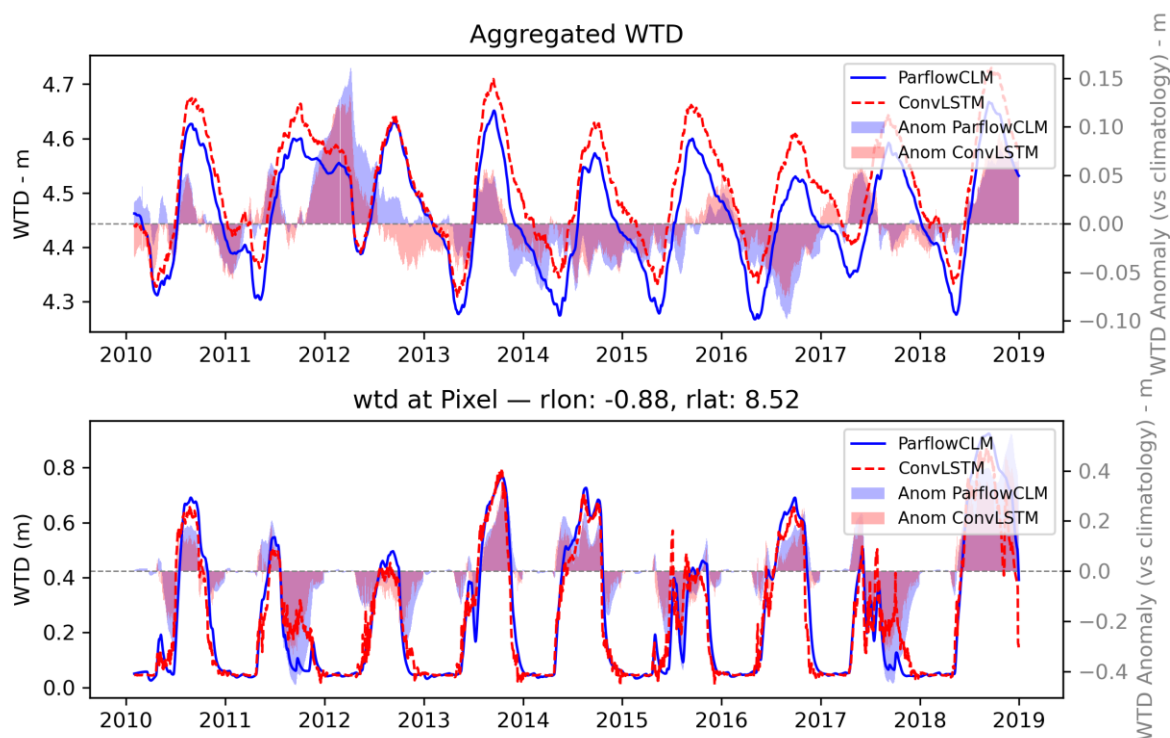


Figure 3.15. Temporal evaluation of the ConvLSTM (driven with ERA5 data) model against the TSMP domain for the period 2010–2019. The upper panel shows the domain-aggregated daily WTD time series, and the lower panel presents the WTD time series at a single representative pixel located.

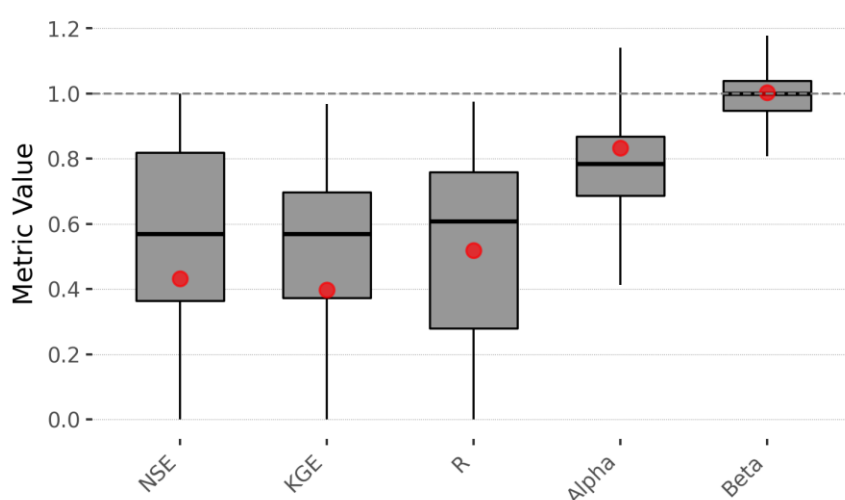


Figure 3.17. Summary performance of the ConvLSTM surrogate forced with ERA5-Land and evaluated against TSMP over the pan-European domain. Boxplots summarize NSE, KGE, Pearson correlation, variability ratio, and bias ratio across valid grid cells, illustrating the degradation in temporal skill and variability representation relative to the TSMP-forced configuration.

A particularly revealing outcome emerges when the ConvLSTM surrogate forced with ERA5-Land is evaluated against independent in-situ groundwater observations rather than against the TSMP reference simulation (Figure 3.18). Despite the degraded performance identified when comparing the ERA5-forced surrogate against TSMP outputs, the spatial correlation maps show that the ConvLSTM-ERA5 configuration achieves broadly comparable and in several stations even superior correlation with observed groundwater anomalies relative to TSMP itself, suggesting that a significant fraction of the discrepancy identified in the previous analysis reflects systematic biases inherent to the TSMP model rather than genuine errors in the surrogate's representation of real groundwater dynamics. This interpretation is supported by the pixel-scale time series, where the ConvLSTM-ERA5 frequently tracks the observed WTD anomalies as closely as or more faithfully than TSMP, correctly reproducing the timing and sign of interannual fluctuations including the progressive drying trend leading at several stations. These results highlight that the operational value of the ConvLSTM surrogate should not be assessed exclusively through its agreement with physically based model outputs, but through its ability to reproduce observed groundwater behavior, a perspective under which the ERA5-forced framework demonstrates meaningful predictive skill even without bias correction, positioning it as a computationally efficient tool for large-scale groundwater monitoring across Europe.

3.5. Conclusions

This chapter presented the development, evaluation, and continental-scale transferability of a ConvLSTM-based surrogate model for estimating daily Water Table Depth, trained on high-resolution ParFlow/CLM DE06 simulations over Central Europe and subsequently transferred, without retraining, to the full pan-European domain using TSMP simulations as an independent benchmark. Unlike the autoregressive SWE framework described in Chapter 2, the WTD model was designed to estimate current groundwater states from antecedent precipitation and soil moisture conditions, using the Antecedent Precipitation Index as a compact representation of recharge memory. The results demonstrate that this approach successfully captures the dominant spatiotemporal patterns of groundwater dynamics across markedly different spatial scales and hydrogeological settings.

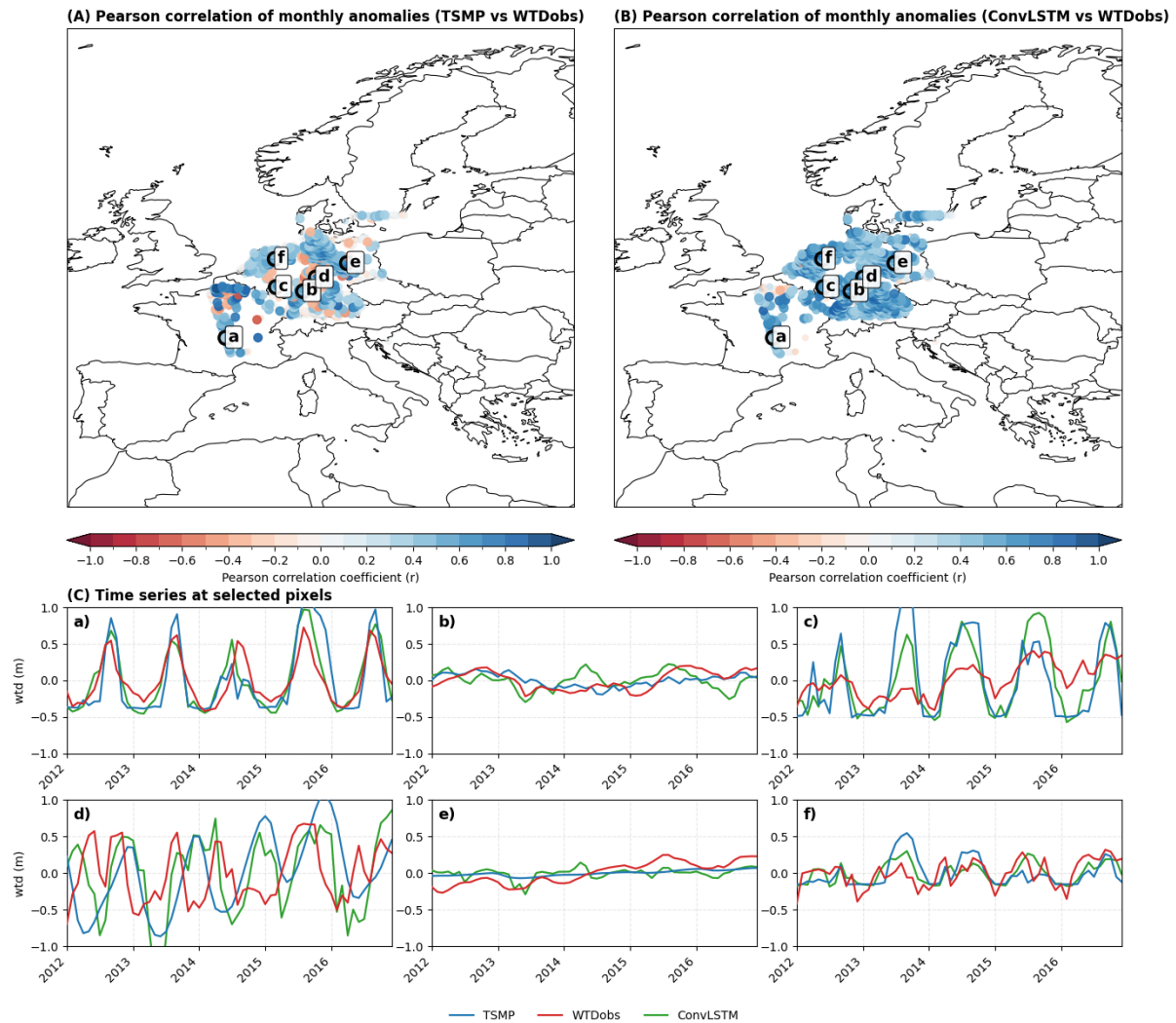


Figure 3.18. Pearson correlation of monthly WTD anomalies between simulated and in-situ observed water table depth in Europe. (A) Spatial distribution of the Pearson correlation coefficient (r) between ParFlow/CLM DE06 and in-situ observations (WTDobs). (B) Spatial distribution of the Pearson correlation coefficient (r) between the ConvLSTM surrogate and WTDobs at the same well locations. (C) Time series of monthly WTD anomalies at six representative stations over the period 2012–2016.

Over the Central European training domain, the ConvLSTM model achieved strong performance when forced with the native HRES-based ParFlow/CLM forcing, with median NSE values near 0.8, KGE near 0.75, and Pearson correlation near 0.9. Skill was highest across the flat northern lowlands of Germany and the Netherlands, where shallow, precipitation-responsive water tables align closely with the recharge signal encoded by the API, while performance degraded systematically toward the Alpine foothills and topographically complex uplands, where lateral subsurface flow and steep hydraulic gradients are less well represented by the model's static and dynamic inputs. When the same architecture was forced with ERA5-Land reanalysis data instead of HRES, performance declined consistently across all metrics (median NSE ~ 0.45 , KGE ~ 0.5 , $r \sim 0.6$) and a pronounced positive bias of +7.5% to +10% emerged across the northern and central portions of the domain, confirming that forcing resolution, rather than model architecture, is the primary factor limiting surrogate accuracy. Despite this degradation, validation against independent in-situ WTD observations showed that the ERA5-Land-driven ConvLSTM retained predictive skill comparable to the ParFlow/CLM reference at most stations, particularly in the northern lowlands, supporting its potential as a computationally efficient alternative for groundwater monitoring even in the absence of high-resolution regional

forcing. The trained model to the full European domain, evaluated against TSMP simulations on the CORDEX EUR-11 grid, represented a substantially more demanding test of generalization, given the pronounced continental-scale heterogeneity in topography, soil properties, and climatic regimes relative to the Central European training environment. Despite the absence of any fine-tuning, the model achieved median NSE and KGE values near 0.8 and Pearson correlation exceeding 0.9, with the strongest skill concentrated across the North European Plain and progressive degradation across the Alps, the Carpathians, and the Balkan ranges, consistent with the structural absence of snowmelt-related processes in the input forcing. A clear latitudinal bias gradient emerged from the Pbias evaluation, with systematic overestimation (+6% to +10%) across Scandinavia and systematic underestimation (−6% to −10%) across the Iberian Peninsula and Mediterranean coastal zones, reflecting the mismatch between the shallow, recharge-dominated groundwater regime of the training domain and the deeper, geologically controlled aquifer systems characteristic of southern Europe. Notably, the model captured the anomalously deep-water tables associated with the 2018 European drought without any domain-specific calibration, underscoring its ability to generalize learned hydrological relationships to non-stationary climatic extremes.

When the pan-European transfer was repeated using uncorrected ERA5-Land forcing in place of the native TSMP meteorological inputs, performance against the TSMP reference degraded substantially (median NSE ~ 0.45 , KGE ~ 0.4 , $r \sim 0.5$), with the most severe deterioration over the British Isles, France, and the Mediterranean coastal regions. However, a particularly significant finding emerged when this same ERA5-Land-driven configuration was evaluated against independent in-situ groundwater observations rather than against TSMP itself: the ConvLSTM surrogate achieved correlation values comparable to, and at several stations exceeding, those of TSMP. This result indicates that a substantial portion of the apparent skill loss reflects systematic biases inherent to the TSMP reference simulation rather than genuine deficiencies in the surrogate's representation of real groundwater behavior, and demonstrates that a ConvLSTM model trained on physically consistent ParFlow/CLM simulations and forced with globally available, bias-uncorrected ERA5-Land data can match or outperform a fully coupled terrestrial system model when judged against ground-truth observations.

Collectively, these results establish that the proposed ConvLSTM framework constitutes a physically grounded, computationally efficient, and genuinely transferable tool for continental-scale groundwater monitoring, capable of generalizing learned spatiotemporal relationships well beyond its original training domain without retraining. The primary remaining limitations relate to the representation of WTD variability amplitude, evidenced by variability ratios persistently below 1.0 across nearly all configurations, and to systematic regional biases in topographically complex or climatically distinct settings such as the Alps and the Mediterranean basin. Future work should prioritize the application of quantile-quantile bias correction to ERA5-Land forcing fields, targeted fine-tuning using local in-situ observations in the Iberian Peninsula and Alps, and the explicit incorporation of snowmelt-related predictors to improve representation of delayed recharge processes in snow-influenced catchments. More broadly, the comparable or superior performance of the surrogate relative to the fully coupled TSMP system when validated against real-world observations highlights the value of evaluating such frameworks not only against physically based reference simulations, but directly against observational ground truth.

4. Predictive mapping of groundwater quality

The MLMapper predictive mapping tool was initially developed and applied within the STARS4Water project to support the prediction of groundwater contamination (Baron & Keller, 2026). Its first application focused on the Duero River basin, where machine learning models were used to integrate multiple environmental and hydrogeological variables and generate spatially predictive maps of nitrate contamination in groundwater. This application demonstrated the potential of MLMapper as a data-driven approach for identifying spatial patterns of groundwater contamination and supporting the assessment of areas potentially at risk. In this report, MLMapper is further applied to the East Anglia region, an area facing significant challenges related to nitrate contamination of groundwater. The objective is to assess the applicability of the tool in a different hydrogeological and environmental context and to generate predictive maps of nitrate contamination. Traditional geostatistical approaches often struggle to account for multiple complex variables simultaneously, and mathematical models require a high degree of discretization of the environment and high-resolution input data, which cannot always be achieved. To overcome this limitation, MLMapper 2.0, a tool developed by the Universidad Complutense Madrid and optimized in partnership with the UK Centre for Ecology and Hydrology, was applied to integrate the available environmental and hydrogeological information into a data-driven predictive mapping framework.

The study conducted in the East Anglia region focused specifically on the Chalk aquifer. These aquifers exhibit clearly unconfined conditions along their western boundaries and confined or semi-confined conditions in the central and eastern areas. In this case, a dataset of 156 monitoring points was used (of which 41% were classified as contaminated). Originally, the database contained a larger number of data points, but after preprocessing and analysis, some were removed because they did not meet certain quality requirements. Based on 2020–2024 nitrate time series (minimum 10 samples per borehole), the classification threshold was set at the legal limit of 50 mg/L, labelling boreholes as "contaminated" if they exceeded this value in more than 10% of samples during this period.

Spatial predictors, such as geological, topographical, soil, climatic, and potential contamination sources, were selected based on scientific literature and stakeholder expertise to model groundwater nitrate contamination. Land use serves as a primary predictor, as agricultural activities and waste discharge represent major pollution sources. Once introduced, nitrate transport is driven by rainfall and irrigation, while its infiltration is regulated by topography and soil composition. Finally, the rock's hydrodynamic properties and the presence of confining layers or chemical conditions directly dictate groundwater flow, either accelerating transport or enabling critical nitrification and denitrification processes.

Table 4.1 presents the predictive performance metrics of the tree-based ensemble models trained using the East Anglia dataset. Rather than directly transferring or fine-tuning the models previously developed for the Duero River basin, the models were trained specifically for East Anglia due to substantial differences in the availability and type of environmental and hydrogeological predictor variables between the two study areas. Model training and evaluation were performed using a repeated nested cross-validation framework, providing robust estimates of predictive performance while reducing the dependence on a single train-test split. All models achieved test scores above 0.83, with Random Forest Classifier (RFC), AdaBoost Classifier (ABC), and Gradient Booster Classifier (GBC) pushing past 0.89 to demonstrate outstanding accuracy. This robustness is further supported by the F1-scores, which consistently exceeded 0.80 for both classes and peaked at 0.96 with GBC, confirming that the models effectively identify both contaminated and uncontaminated zones. Finally, the area

under curve (AUC) values ranging from 0.96 to 1.00 represent the highest tier of classification performance, validating the overall reliability of the system.

Table 4.1. Summary of machine learning metrics for the tree-based ensemble models in the East Anglian chalk aquifer. RFC = Random Forest Classifier; ABC = AdaBoost Classifier; GBC = Gradient Booster Classifier; ETC = Extra Trees Classifier. Performance metrics: Train score = optimised training score; Test score = optimised test score; Prec 0 = precision negative class; Prec 1 = precis; F1 0 = F1 score negative class; F1 1 = F1 score positive class; AUC = area under curve.

Algorithm	Train score	Test score	Prec 0	Prec 1	Rec 0	Rec 1	F1 0	F1 1	AUC
RFC	0.98	0.91	0.86	1.00	1.00	0.83	0.92	0.90	0.99
ABC	0.99	0.89	0.85	0.95	0.96	0.83	0.90	0.88	0.97
GBC	1.00	0.96	0.96	0.96	0.96	0.96	0.96	0.96	1.00
ETC	0.97	0.83	0.77	0.94	0.96	0.70	0.85	0.80	0.96

The spatial distribution of the mean probability of exceeding the 50 mg/L nitrate threshold—as predicted by the four ensemble algorithms—is displayed Figure 4.1. The highest probabilities are concentrated across the western margin of the Chalk aquifer, where the formation outcrops directly at the surface. Conversely, the central sector exhibits lower probabilities because the aquifer is confined by Quaternary deposits that act as a protective barrier against surficial nitrate inputs. Toward the eastern band, surface streams slice through these Quaternary layers, re-exposing the underlying chalk materials and driving up contamination risks. Overall, the models can accurately distinguish between contaminated and uncontaminated boreholes even when they are in close geographical proximity.

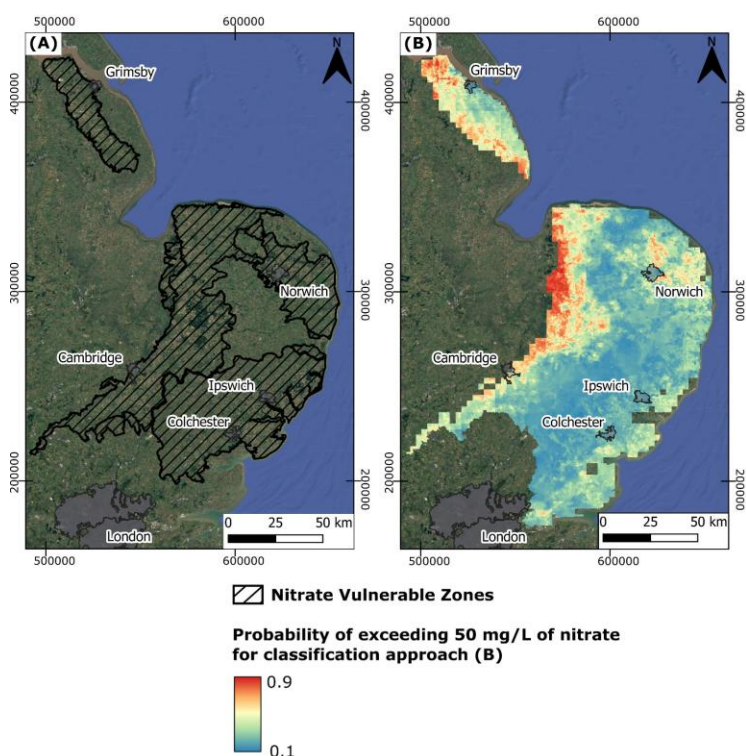


Figure 4.1. Predicted probability of groundwater nitrate contamination in the East Anglia chalk aquifer, expressed as the ensemble mean probability of exceeding the 50 mg/L nitrate threshold. The map also shows Nitrate Vulnerable Zones and highlights higher contamination risk along Chalk outcrop areas and incised valleys where the aquifer is more exposed to surface-derived nitrate inputs.

Figure 4.2 breaks down the underlying factors that determine nitrate contamination in the study locations to analyse the inner mechanisms of the top-performing GBC classifier. Figure 4.2 shows a bee-swarm plot generated using SHapley Additive exPlanations (SHAP) (Nohara et al., 2022). This technique ranks predictor variables by importance from highest to lowest, while showing whether high or low values of a specific factor increase or decrease the likelihood of contamination. According to this analysis, the thickness of the Quaternary material and the Chalk base elevation emerge as the primary controlling factors. Specifically, higher values of Quaternary thickness have a negative impact on the predictions—indicating lower contamination risks—while a higher aquifer base elevation correlates with a higher probability of contamination.

The resulting map is consistent with both the conceptual groundwater model and the available field data. Notably, it effectively represents the presence of nitrate-vulnerable zones along the outcrops of Chalk aquifer. A clear contrast can be observed between these zones and areas where productive aquifers are shielded by overlying impermeable or semi-impermeable layers. Additionally, the map successfully highlights deeply incised valleys on the eastern side of the basin, where conditions resemble those found in the Chalk outcrop areas. The experience of successfully applying MLMapper in another European basin demonstrates its high potential for extrapolation, if a reliable network of existing boreholes is available.

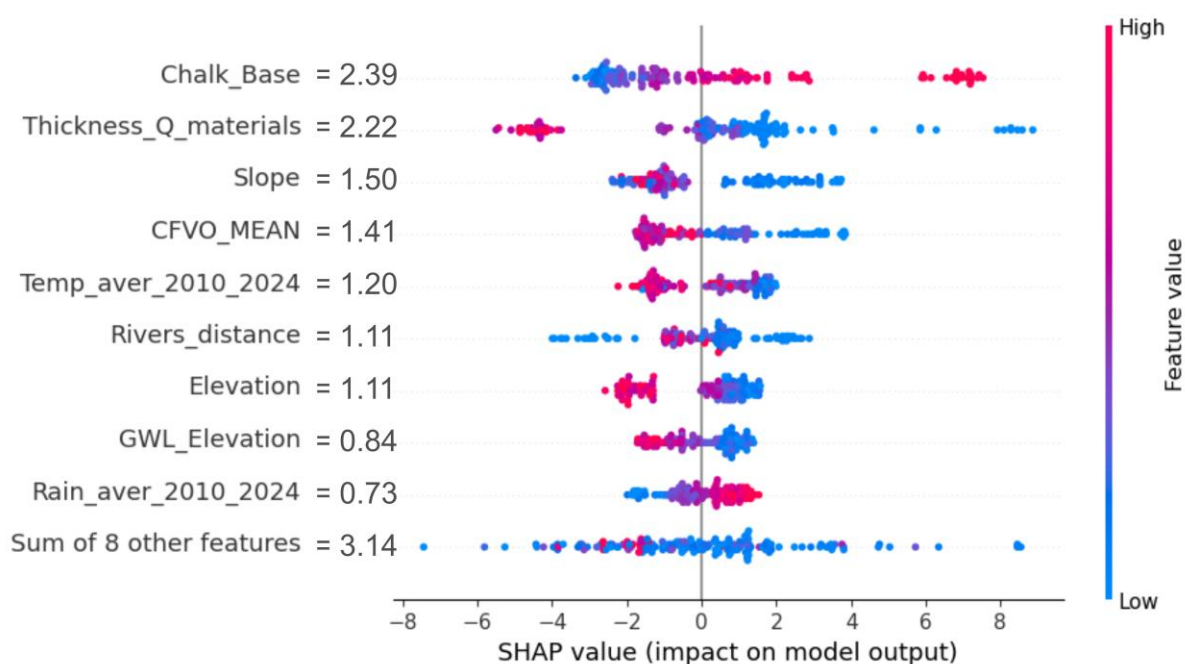


Figure 4.2. SHAP bee-swarm plot for the Gradient Boosting Classifier. The numbers displayed next to each predictor correspond to their mean absolute SHAP values, which determine the ranking of variable importance. The plot illustrates both the relative influence of each predictor and how their intrinsic values contribute to increasing or decreasing the predicted probability of exceeding the nitrate threshold.

5. Ensemble-tree models for reservoir storage

One of the data driven approaches developed within STARS4Water was 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. The aim of this model is to help improve basin-wide assessments of water resources at seasonal to sub-seasonal timescales, an area identified by stakeholders. The model performed well, with a median NSE score of 0.91 for one-month ahead predictions, and 0.8 for three-month ahead predictions. More details on model development and performance can be found in Baron & Keller (2026).

This chapter presents the results of an application of the model to four reservoirs in Seine river basin, with and without model retraining, to evaluate how transferable the model is for reservoirs in European river basins. The reservoirs studied are detailed in Table 5.1. Three reservoirs are non-impounding (i.e. are primarily filled through pumped water) and one is impounding (i.e. reservoirs formed by damming a natural watercourse). The non-impounding reservoirs are operated for water supply and flood control, which requires careful balancing of the two purposes to ensure that there is remaining sufficient storage capacity in the reservoir during high flows to reduce flood risk while maintaining stocks for water demand later in the season. These reservoirs are quite different to the original training set of reservoirs, which are primarily impounding reservoirs, and primarily used for irrigation and water supply. Some are dual-purpose reservoirs where hydropower is the additional purpose (for full details, see Table 7 in Baron & Keller, 2026).

Table 5.1. Details of the Seine reservoirs (WS = water supply; P = hydropower, FC = flood control).

Reservoir Name	Basin	Impounding or Non-Impounding	Purpose ¹	Catchment Area (km ²)	Capacity (Mm ³)
Lac de Pannecière	Seine	I	WS, P	220	82,500
Lac d'Orient	Seine	N-I	WS, FC	46	217,000
Lac du Der-Chantecoq	Seine	N-I	WS, FC	90	364,500
Lac du Temple	Seine	N-I	WS, FC	56	162,500

The ET model for forecasting reservoir storage was run for 1- and 3-month ahead predictions for the Seine reservoirs, with and without model retraining. Model skill was evaluated for the period 2014 to 2021, using the Nash-Sutcliffe Efficiency (NSE), Root Mean Square Error (RMSE) and Percentage Bias (Pbias). The optimal value for these metrics is 1, 0 and 0 respectively. The NSE score was compared to a simple benchmark model (SARIMAX, for more information see Baron et al., 2026), and the change in skill (Δ NSE) was calculated. A positive value of Δ NSE indicated improved performance compared to the benchmark, with a negative value indicating reduced performance.

The results are presented in Table 5.2 (1 month ahead) and Table 5.3 (3 months ahead), with some example time-series presented in Figure 5.1. These results show a clear improvement in ET model performance with retraining. Although the skill metrics show reasonable performance for the untrained predictions (NSE of around 0.9 for 1 month ahead), these are due to the highly predictable nature of reservoir storage (particularly for these highly controlled reservoirs).

The single impounding reservoir, Lac de Pannecière, generally shows high performance and higher change in skill compared to the non-impounding reservoirs.

D4.4 FUTURE OF EUROPE'S FRESHWATER RESOURCES

Table 5.2. Summary of results for 1-month ahead storage predictions without model retraining (untrained) and with model retraining (trained), evaluated using Nash-Sutcliffe Efficiency (NSE), Root Mean Square Error (RMSE), Percentage Bias (Pbias), and change in skill Δ NSE.

Reservoir name	untrained				trained			
	NSE	RMSE (hm ³)	Pbias (%)	Δ NSE	NSE	RMSE (hm ³)	Pbias (%)	Δ NSE
Lac de Pannecière	0.91	7.7	5.1	-0.96	0.98	4.1	-0.7	0.43
Lac d'Orient	0.90	8.8	2.8	-0.86	0.96	5.4	-1.0	0.31
Lac du Der-Chantecoq	0.90	8.6	3.1	-2.45	0.97	4.5	-0.4	0.04
Lac du Temple	0.90	8.0	2.3	-1.00	0.96	5.0	-2.0	0.22

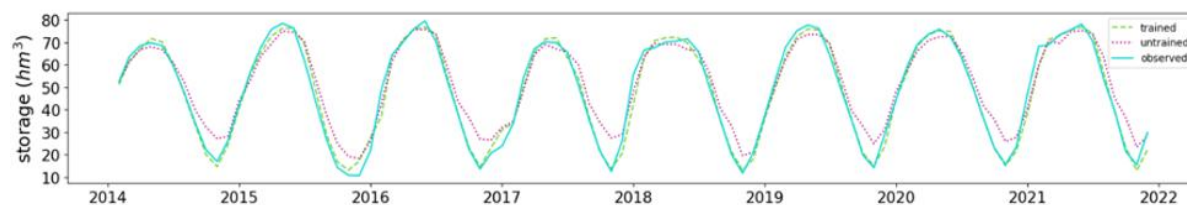
Table 5.3. Summary of results for 3-month ahead storage predictions without model retraining (untrained) and with model retraining (trained), evaluated using Nash-Sutcliffe Efficiency (NSE), Root Mean Square Error (RMSE), Percentage Bias (Pbias), and change in skill Δ NSE.

Reservoir name	untrained				trained			
	NSE	RMSE (hm ³)	Pbias (%)	Δ NSE	NSE	RMSE (hm ³)	Pbias (%)	Δ NSE
Lac de Pannecière	0.50	18.9	15.7	-1.37	0.94	6.3	-0.9	0.74
Lac d'Orient	0.47	20.3	7.9	-4.34	0.93	7.6	-1.8	0.26
Lac du Der-Chantecoq	0.46	20.4	9.8	-8.11	0.94	6.6	-0.5	0.02
Lac du Temple	0.51	18.7	7.2	-3.23	0.89	8.4	-3.5	0.07

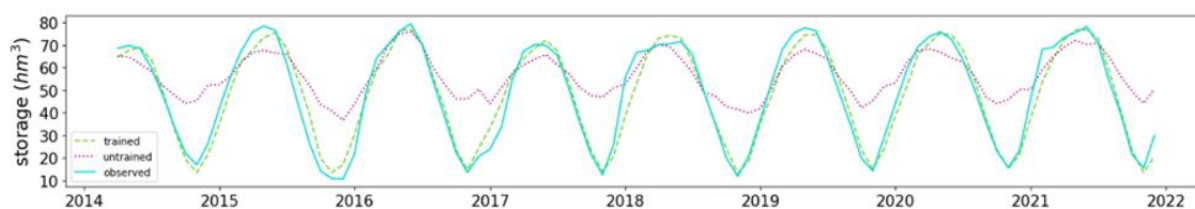
Comparing these to the performance of a simple benchmark model there is no value in using the ET model for these reservoirs with no retraining, at either lead time. With retraining, the ET model shows an increase in predictive skill compared to the simple benchmark for all reservoirs at both lead times, although with limited improvement for Lac du Der-Chantecoq (due to the high NSE value for the benchmark at that reservoir).

While the ET model can be easily applied outside of the training region, these results suggest that there is limited benefit in doing so when the new reservoirs have significant differences to those seen at training time. Retraining the model with the additional reservoirs gives a clear improvement in predictive skill. While these results are suggestive, a detailed leave-one-out analysis would be required to make more concrete statements.

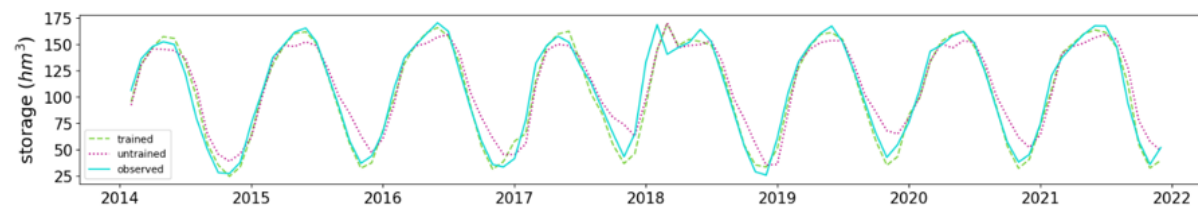
Lac de Pannecière, 1 month ahead



Lac de Pannecière, 3 months ahead



Lac du Temple, 1 month ahead



Lac du Temple, 3 months ahead

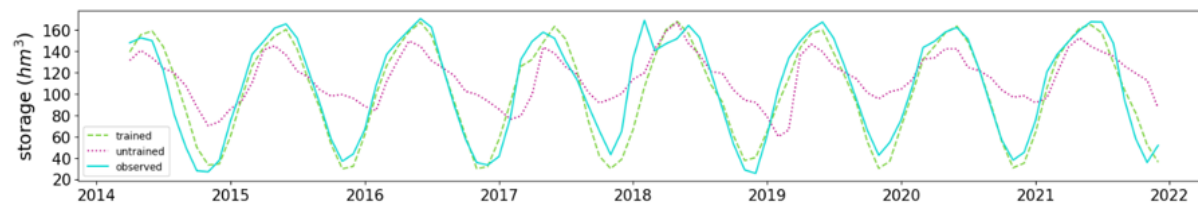


Figure 5.1. Monthly time series of observed and predicted reservoir storage for Lac de Pannecière and Lac du Temple at one- and three-month forecast horizons. Panels compare observations with ensemble-tree model predictions before and after local retraining, illustrating the improvement in forecast skill after adapting the model to the Seine reservoir system.

6. Ensemble LSTM transfer learning for Water Table Depth

Climate change and increasing water stress underscore the need for reliable hydrological modelling. However, the uneven distribution of hydrometeorological observations limits process understanding and hinders the development of data-driven models, particularly in data-scarce regions (Fekete & Vörösmarty, 2007; Do et al., 2017). Groundwater, and particularly Water Table Depth (WTD), is a key variable linking subsurface and land-surface processes, making its accurate estimation necessary for sustainable water management (Kollet & Maxwell, 2008; Wang et al., 2014). Recent studies have demonstrated the ability of Long Short-Term Memory (LSTM) networks to capture groundwater dynamics across diverse environmental conditions (Zhang et al., 2018; Liang et al., 2021; Vu et al., 2021; Taranto et al., 2025). Transfer learning further extends the applicability of such models to data-limited regions by leveraging information learned from data-rich domains, with several hydrological studies reporting improved generalization and predictive performance across watersheds and continental scales (Kratzert et al., 2019; Ma et al., 2021; Li et al., 2021; Xu et al., 2023; Ma et al., 2024; Ryd and Nearing, 2025). Nevertheless, assessing prediction skill in regions lacking observations remains a challenge. Ensemble-based uncertainty quantification provides a practical solution, as the spread among ensemble members can serve as an indicator of prediction confidence (Buizza, 1997; Buizza, 2006; Chapman et al., 2022; Hu et al., 2023).

Although transfer learning provides a promising pathway for large-scale WTD monitoring, assessing prediction quality in regions without observations remains challenging. To address this limitation, this study develops a scalable and transferable WTD prediction framework based on an ensemble of LSTM networks trained across the European-Coordinated Regional Experiment (EURO-CORDEX) domain using readily available hydrometeorological inputs (Giorgi et al., 2009; Gutowski et al., 2016). Model transferability is evaluated on grid cells excluded from training following an approach similar to that of Kratzert et al. (2019), while prediction reliability in data-scarce regions is assessed through spread–skill relationships that link ensemble spread to model performance.

The study uses the Terrestrial Systems Modelling Platform Ground to Atmosphere (TSMP-G2A) dataset, which provides hourly hydro-climatic simulations across continental Europe at a spatial resolution of 0.11 (Furusho-Percot et al., 2019). The underlying TSMP-G2A framework combines the atmospheric model COSMO, the land-surface model CLM, and the groundwater model ParFlow. An ensemble of LSTM networks is trained to predict daily WTD using dynamic hydrometeorological variables and static attributes derived from the dataset. Training is conducted for the period 2002–2015, while model evaluation covers 2017–2020. Ensemble learning is achieved through bootstrap sampling of spatially distributed grid cells, with each sample used to train an individual LSTM model.

The grid cells located at the domain north, east, and west (~125 km) and south (~1250 km) boundaries of the EURO-CORDEX domain were excluded to minimize boundary effects and remove non-relevant regions. In addition, grid cells exhibiting constant WTD values, identified by zero variance over a six-month period, were removed since the objective is to predict dynamic groundwater behaviour. After applying these filtering criteria, Figure 6.1 illustrates the remaining 43,226 valid grid cells for model development and analysis within the EURO-CORDEX domain.

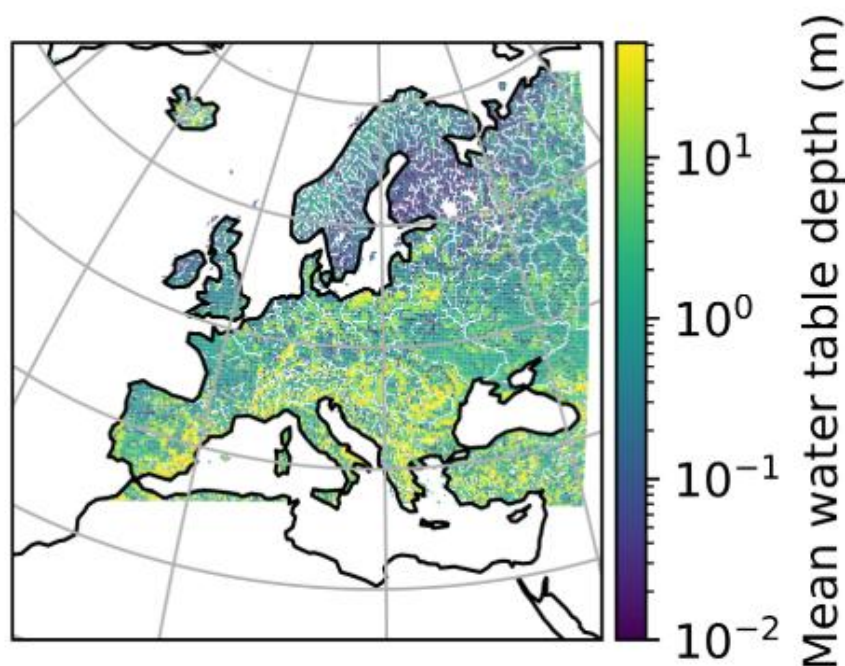


Figure 6.1. Mean water table depth (WTD; m) across the EURO-CORDEX domain in 2020. Only grid cells satisfying the applied filtering criteria are shown and included in the subsequent analysis.

The model predicts WTD using an LSTM network that integrates dynamic hydrometeorological inputs with static attributes. Temporal dependencies are learned through a lookback window of 365 days, allowing the model to capture time-lagged relationships in groundwater dynamics. Input variables include precipitation, vapor pressure deficit, and soil moisture at 1 m depth, all of which are spatially and temporally continuous across Europe and can be derived from observations or reanalysis products such as ERA5-Land. Precipitation was aggregated to the daily sum, while vapor pressure deficit, soil moisture, and WTD were aggregated to the daily mean. Spatial variability is incorporated through static attributes (longitude, latitude, slope in longitudinal and latitudinal directions, and soil type), which remain constant over time for each grid cell. All input variables are aggregated to a daily resolution to match the temporal scale of WTD, which evolves over long timescales.

The LSTM architecture is configured with 256 hidden units and produces a single-step forecast of WTD at time t . The model is trained for 100 epochs using the Adam optimizer (Kingma & Ba, 2014), with an initial learning rate of 0.001 until epoch 50, followed by a reduced rate of 0.0001. Mean squared error is adopted as the loss function. Input data are structured as a three-dimensional tensor consisting of batch size, lookback length, and number of features. Although the original dataset spans temporal and spatial dimensions, these are combined into a unified sample dimension so that each training instance corresponds to a 3D input sequence suitable for LSTM processing.

Ensemble learning is implemented using the pasting approach, i.e., bootstrapping without replacement (Breiman, 1999), where each ensemble member consists of an independent LSTM network. The size of the bootstrap sample (n), defined by the number of grid cells, controls the amount of spatial information available to each model. To investigate the impact of increasing spatial coverage, two experimental setups are designed by varying n from 100 to 400 grid cells. For each experiment, 100 ensemble members are trained, a choice based on achieving a balance between predictive accuracy, ensemble diversity, variance reduction, and computational cost. Consequently, the two experiments use 10,000 and 40,000 grid cells randomly sampled across the study domain for

training (the trained subsets), while the remaining 3,226 grid cells, which are consequently also randomly distributed across the domain, are unseen during training and reserved for transfer (the transfer subset). Table 6.2 summarizes the distribution of training and transfer subsets. Training a single LSTM model required approximately 75 minutes on four NVIDIA A100 GPUs.

Table 6.2. Dataset, training, and transfer subset grid cell count.

Dataset	Count (grid cell)
Full dataset	43,226
Trained subset	10,000/40,000
n (number of a single LSTM training grid cells)	100/400
Transfer subset	3,226

The model was trained over the period from 2002 to 2015, while the testing is assessed over the test period from 2017 to 2020 on the training subset, by evaluating each LSTM on its respective training grid cells over the test period and comparing predicted WTD against the TSMP-G2A reference dataset. Performance is quantified using Pearson's correlation coefficient, Root Mean Square Error (RMSE), Kling–Gupta Efficiency (KGE), and Mean Bias. Transfer learning is evaluated over the same test period. In a transfer setup, predictions are obtained by averaging the outputs of all 100 ensemble members in the transfer subset, whereas for the trained subset, the mean of 99 members is used for each particular grid cell, excluding the member trained on the respective grid cell, under the assumption that excluding a single member has a negligible effect on the ensemble mean. This design is consistent with the bootstrapping-without-replacement strategy to exclude, in a transfer setup, a single member that was trained on a certain grid cell and ensures comparability between trained and transfer evaluations.

Figure 6.2. summarizes the training, testing, and transfer workflows. In addition, WTD predictions over the transfer subset are used to relate ensemble spread to model performance. Prediction quality on the transfer subset is determined from ensemble spread by relating it to forecast performance metrics. This is achieved through spread–skill relationships that link uncertainty indicators to model accuracy measures. We focus in this work on RMSE and mean bias, such that the coefficient of determination (R^2) is considered as the goodness-of-fit criterion. Ensemble spread is derived from the predictions on the transfer subset, produced by ensemble members over the testing period. The ensemble uncertainty is quantified from ensemble spread: ensemble variance (EV), defined as the variance across members at each time step and then averaged over time, and the interquartile range (IQR), computed between the 25th and 75th percentiles at each time step and subsequently averaged over the full time series. These spread measures are then related to the performance metrics through the proposed spread–skill relationships to evaluate whether ensemble uncertainty can serve as a proxy for predictive accuracy in data-scarce regions, a similar approach to the Ensemble Prediction Systems (EPS) (Buizza, 1997; Buizza et al., 2006).

Both trained LSTM ensembles (based on $n=100$ and $n=400$) were evaluated on the testing and transfer grid cells over the evaluation period to assess the benefits of increasing the amount of spatial training information. The cumulative distribution functions of RMSE, Pearson correlation, Mean Bias, and KGE are presented in Figure 6.3. Increasing the number of training grid cells from 100 to 400 per ensemble member improved model performance, demonstrating enhanced generalization and transferability. In the transfer subset, the 400 gridcell-based ensemble achieved global improvements of 11.5% in RMSE and 17.4% in Pearson correlation, while maintaining a similar mean bias. These findings indicate

that increasing spatial information and diversity during training enhances predictive performance, consistent with previous studies highlighting the benefits of training on broader spatial domains (Ma et al., 2024; Kratzert et al., 2024). Consequently, only the $n = 400$ configuration was considered for the subsequent analyses.

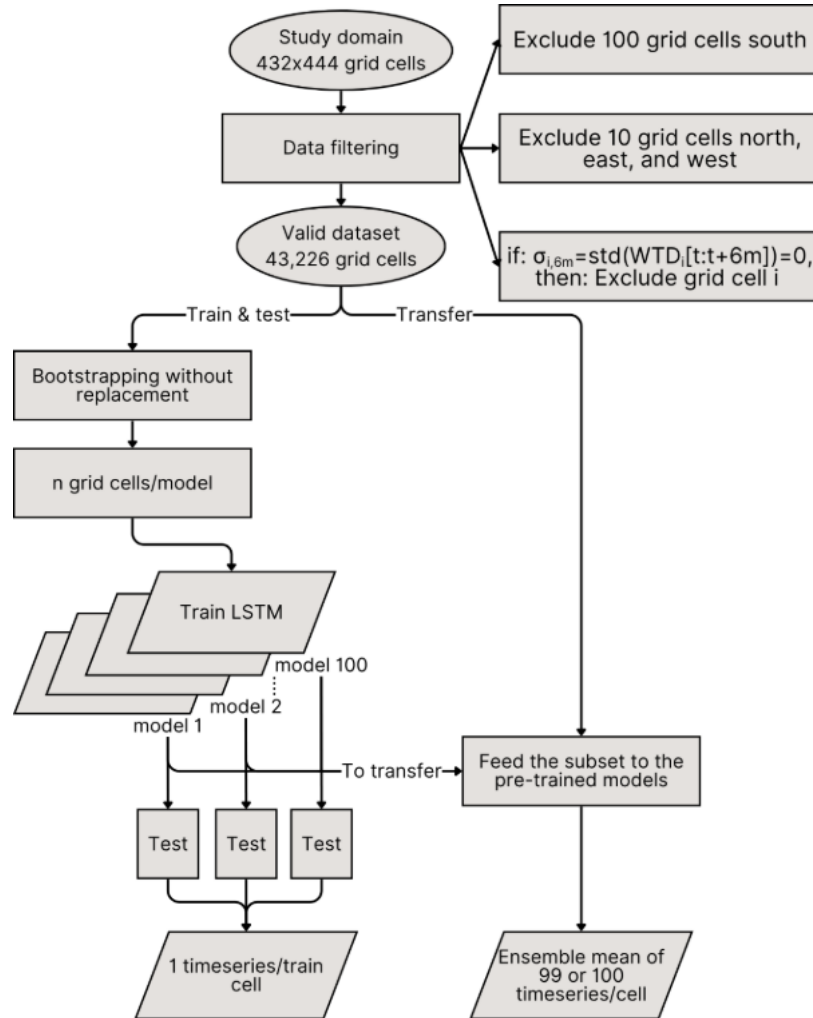


Figure 6.2. Workflow for data filtering and bootstrap sampling, integrated with the LSTM ensemble training, testing, and transfer procedures.

Although transfer performance was lower than testing performance, the ensemble maintained strong predictive skill on unseen locations. Pearson correlation remained nearly unchanged between testing and transfer, indicating that the model successfully reproduced temporal WTD dynamics. However, transfer predictions exhibited a higher RMSE, where about 90% of grid cells having RMSE values below 1 m during testing, compared to only 50% during transfer, suggesting reduced accuracy in estimating the magnitude of groundwater fluctuations. KGE values were generally lower, such that, for $n=400$, approximately 40% of grid cells show a KGE greater than 0.6 in testing, compared to only 15% in transfer, particularly for low-variance time series, where the variability component strongly influences the metric, which can lead to poor scores (Knoben, 2024). Overall, the results demonstrate that the ensemble effectively captures temporal groundwater behaviour in unobserved locations.

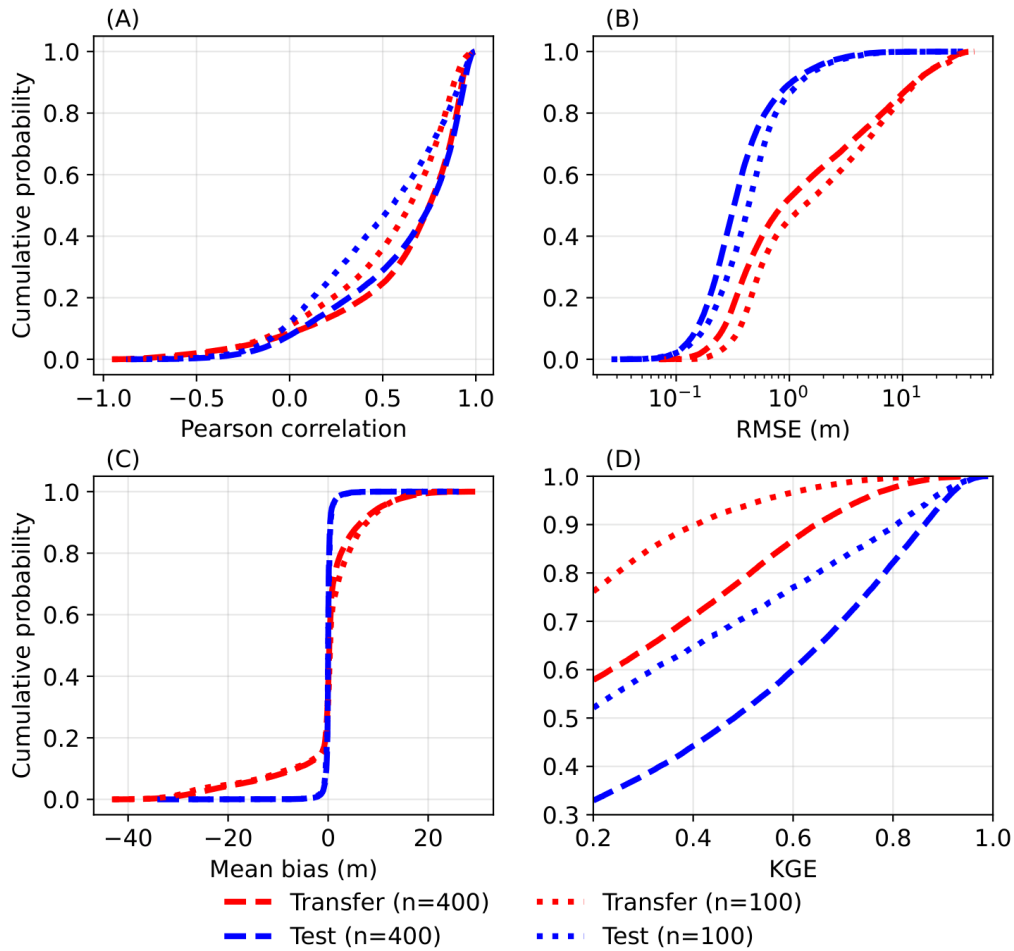


Figure 6.3. Cumulative distribution functions of the performance metrics (A) Pearson correlation, (B) RMSE, (C) Mean bias, and (D) KGE of the ensemble mean versus observations for $n=100$ and $n=400$ ensembles.

A key challenge in transfer-learning applications is the absence of observations in target regions, which prevents direct evaluation of prediction accuracy. To address this limitation, ensemble spread is used as a proxy to determine model performance by establishing relationships between ensemble spread measures and evaluation metrics focusing only on the transfer subset, not considering the training subset. The derived relationships and the estimated performance from ensemble spread are then compared with observation-based metrics on the full study domain. The results show strong spread–skill relationships for RMSE and absolute mean bias (Figure 6.4), with power-law models achieving coefficients of determination (R^2) greater than 0.75, indicating high predictive capability. These findings suggest that ensemble spread can provide reliable estimates of magnitude-based performance metrics, particularly RMSE and bias, in regions where observations are unavailable.

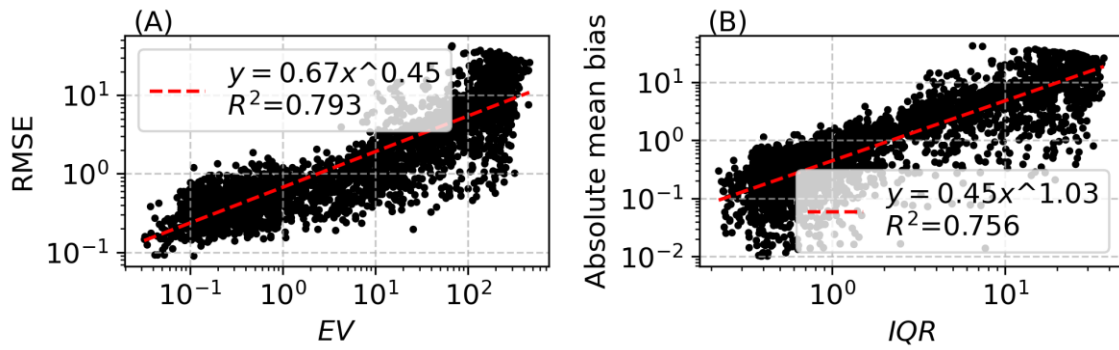


Figure 6.4. Relationships between performance metrics (RMSE and Absolute Mean Bias) and ensemble spread measures Ensemble Variance (A; EV) and Inter Quartile Range (B; IQR).

The developed spread–skill relationships enable ensemble spread to reliably estimate RMSE across the full study domain (training and transfer subsets), producing predictions that closely match observation-based performance metrics as illustrated in Figure 6.5. These results demonstrate the potential of ensemble uncertainty to assess prediction quality in regions without observations, thereby supporting confidence assessment and defining the applicability of transfer-learning predictions in data-scarce areas.

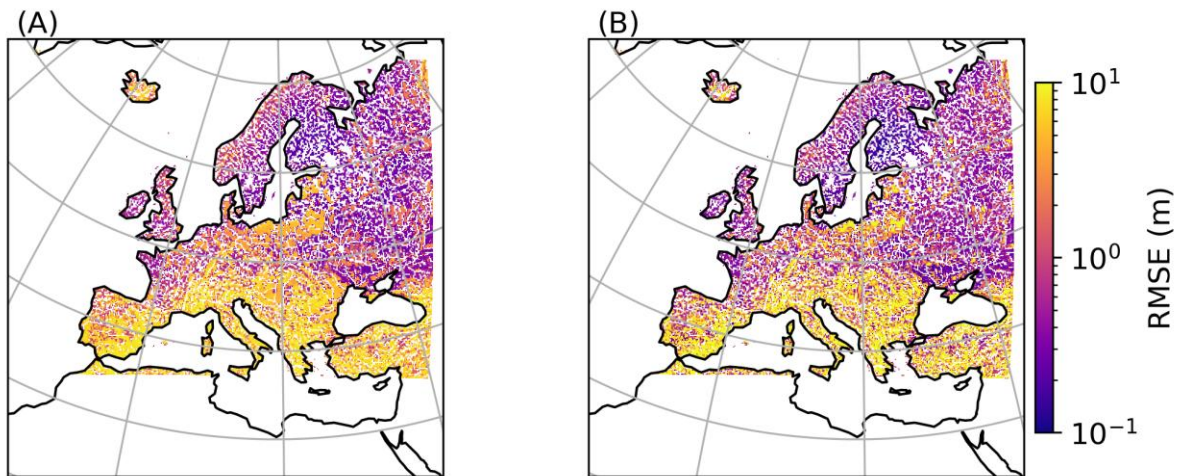


Figure 6.5. RMSE (A) predicted from ensemble variance and (B) evaluated from observations across the full study domain in a transfer setup.

The proposed framework has several limitations. It was developed and evaluated only over the European domain, limiting its applicability to tropical and arid climates. Furthermore, the use of geographic coordinates as input features may reduce model transferability beyond the training region. The analysis relies exclusively on physically based simulations from the TSMP-G2A dataset rather than observational data, and coastal and oceanic processes are not represented.

This study demonstrates the potential of ensemble LSTM transfer learning for large-scale prediction of water table depth using hydrometeorological variables. Training in spatially diverse regions enabled model generalization and transfer to previously unseen locations, reducing reliance on site-specific calibration. Increasing the training subset from 100 to 400 grid cells per ensemble member improved transfer performance by 11.52% in RMSE and 17.42% in Pearson correlation, while maintaining strong

temporal prediction skill in a transfer setup despite some loss in magnitude accuracy. Furthermore, ensemble spread proved to be an effective indicator of prediction quality in regions without observations, particularly for estimating RMSE and absolute mean bias through spread–skill relationships. These findings highlight the value of combining transfer learning with ensemble-based uncertainty quantification to support groundwater monitoring in data-scarce regions and provide a practical framework for assessing prediction confidence where reference observations are unavailable. Future work should extend the framework to global-scale applications and validate its performance against observational groundwater datasets.

7. Groundwater storage using machine learning models

The groundwater storage model originally developed for the Duero River basin (see Baron & Keller, 2026) was transferred to the Seine River basin to assess its ability to reproduce the spatio-temporal dynamics of terrestrial water storage in a contrasting hydrological setting. The model is a hybrid spatio-temporal transformer (STT) deep-learning architecture that predicts monthly variations in total groundwater storage using a combination of static catchment descriptors, dynamic climatic and land-cover drivers, and the recent history of the target variable itself. Because poor correlations were found between observed water table depths in wells and the values simulated by the Terrestrial Systems Modelling Platform (TSMP), the target variable for the Seine application was instead derived from TSMP monthly total water storage fields (ΔTWS), computed as the difference between consecutive monthly time steps. Total Water Storage (TWS) in TSMP is the total amount of water stored from the surface to the deep groundwater in each pixel. Although it is not the exact groundwater storage, it is correlated with it, and it can be used as a proxy to test whether the architecture of the DL model emulator for the Duero can be transferred to assess water storage in a different basin/context.

The model domain covers the Seine basin on a regular 0.11° grid (48×28 cells), of which 741 cells fall within the catchment. The analysis spans 131 monthly time steps between February 2010 and December 2020. Three groups of predictors were used: (i) static descriptors – soil texture and organic carbon from SoilGrids, lithology, surface water bodies and river-network density; (ii) dynamic exogenous drivers – precipitation, air temperature, CORINE land-cover classes and gridded population (GHS-POP); and (iii) the autoregressive history of ΔTWS . Unlike the original Duero implementation, an uncertainty layer was not available for the Seine and was therefore removed, with the loss function reverting to uniform weighting.

The core STT model embeds the static, dynamic and autoregressive inputs and combines the exogenous and autoregressive streams through a learnable gated fusion before a transformer encoder with attention pooling; a spatial embedding of the cell coordinates is concatenated prior to the prediction head, which produces a 12-month forecast horizon. Training samples were generated with a sliding-window scheme (48-month history $\rightarrow$ 12-month prediction), which substantially increased the number of training examples (around 37,000) and mitigated over-fitting. The 741 basin cells were randomly split into training, validation and test sets (70/15/15) so that the test set measures generalisation to unseen locations across the fulltime span. Using the complete ΔTWS series, the autoregressive model clearly outperformed the two climatological baselines (Table 7.1). On the held-out test cells it reached a coefficient of determination of $R^2 = 0.80$ and a mean absolute error of 14.3 mm. The narrow gap between training ($R^2 = 0.89$) and test ($R^2 = 0.80$) performance indicates a well-generalised model with no substantial over-fitting.

Table 7.1. Performance of the groundwater storage models transferred to the Seine basin, evaluated on held-out test cells (selected model in bold). Climatology is a per-cell monthly mean computed only from months preceding each forecast; seasonal-naïve repeats the value observed twelve months earlier. R^2 = coefficient of determination; MAE = mean absolute error of the monthly total water storage variation (ΔTWS).

Model	Train R^2	Val R^2	Test R^2	Test MAE (mm)
Climatology	-	-	0.41	-
Seasonal Naïve	-	-	0.01	-
STT (selected)	0.893	0.715	0.795	14.3

Figure 7.1 shows the times series prediction diagnostic for the Seine. At the three best-fitted cells ($R^2 = 0.94$) the model reproduces both the amplitude and the phase of the monthly signal almost exactly. In the two worst-fitted cells ($R^2 = -0.16$ and -0.31) the prediction collapses towards a nearly flat series and does not capture the observed month-to-month excursions at all, which points to local dynamics that the predictor set does not represent. As in the Duero, such cells are rare: only 2 of the 112 test cells (1.8 %) have a negative R^2 , and the median per-cell value is 0.88 (Figure 7.2).

STT Model — Observed vs Predicted ΔTWS (Seine)

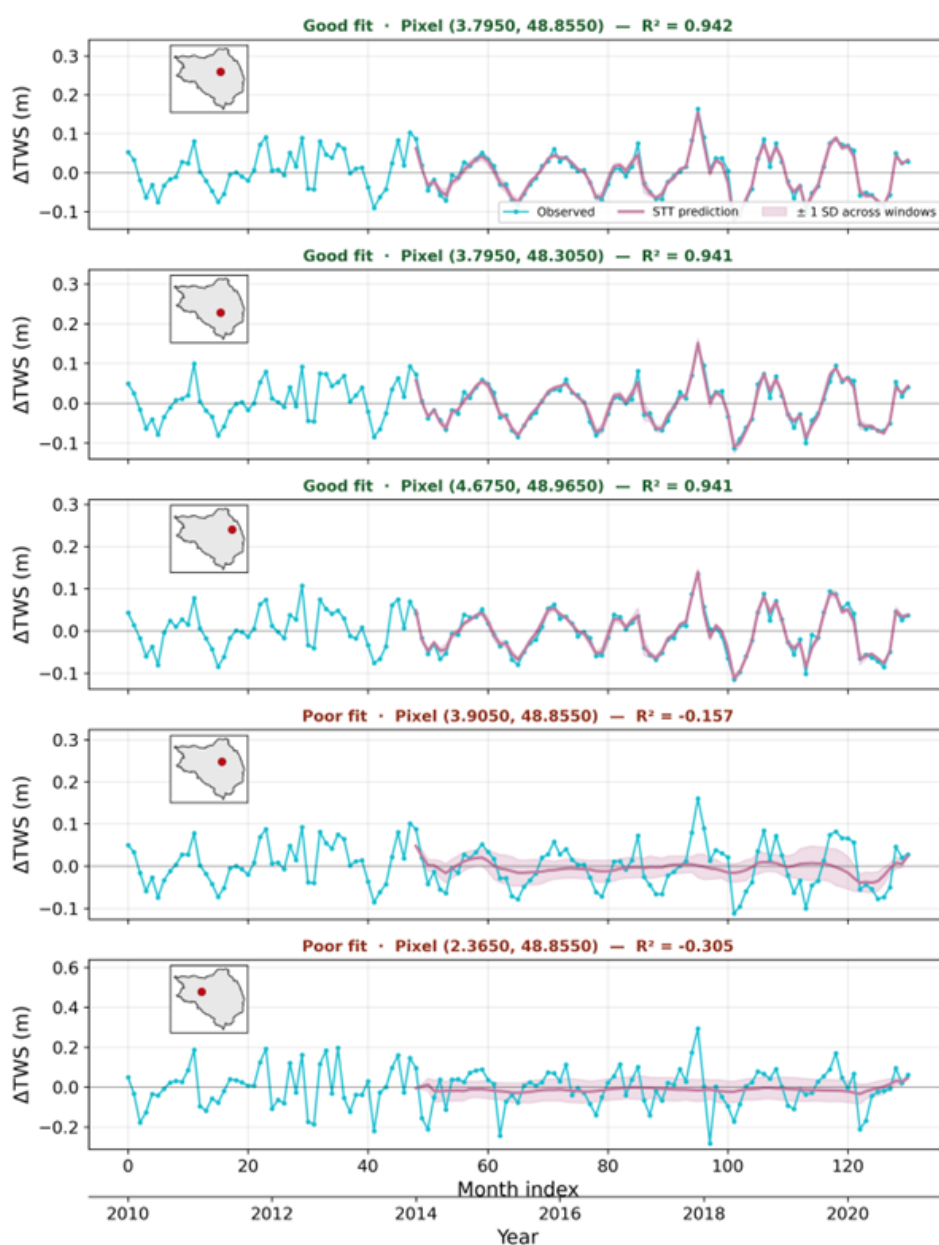


Figure 7.1. Observed TSMP monthly total water storage variation (ΔTWS , cyan) compared with predictions (rose) of the selected spatio-temporal transformer (STT) model at five held-out test cells in the Seine basin: the three with the highest per-cell R^2 (green titles) and the two with the lowest (red titles). The rose line is the mean prediction across all sliding windows covering each month and the shaded band the ± 1 standard deviation. The inset map in each panel locates the corresponding cell (red dot) within the Seine.

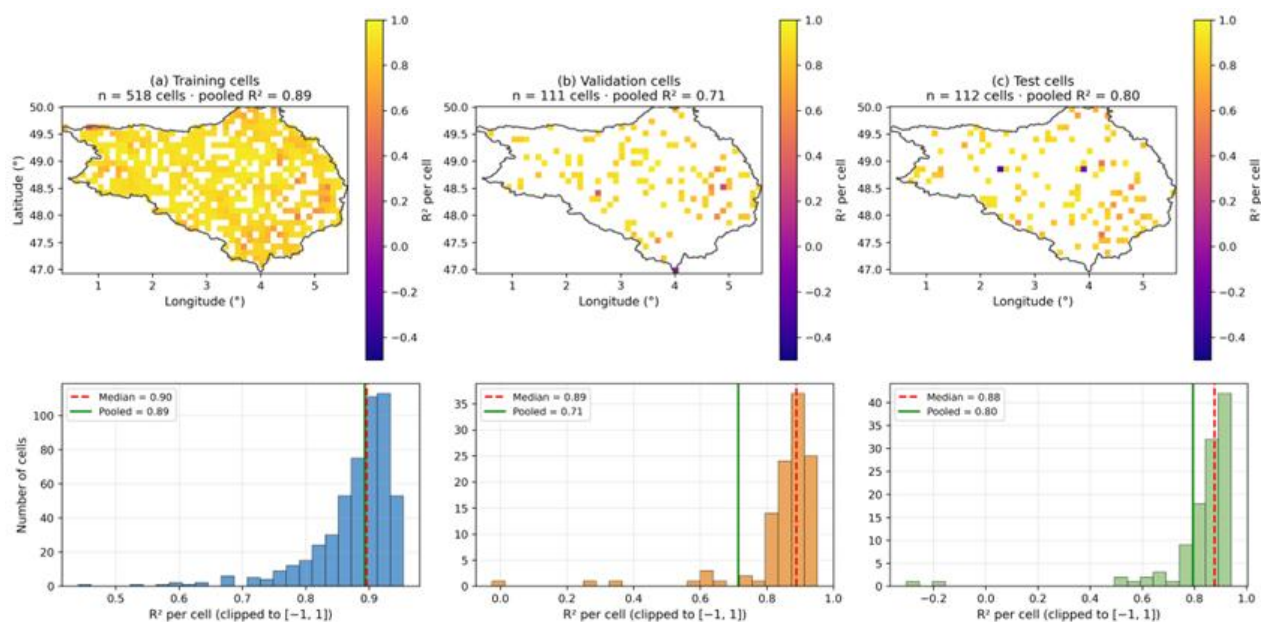


Figure 7.2. Spatial distribution of the per-cell coefficient of determination (R^2) for the selected spatio-temporal transformer (STT) model over the Seine basin, for the (a) training, (b) validation and (c) test cells of the spatial split, with the corresponding histograms below. Sample size and pooled R^2 are annotated on each panel.

A permutation-based analysis of predictor importance identified precipitation, the autoregressive Δ TWS history and air temperature as the three dominant drivers, an order of magnitude above population, land-cover and static descriptors (Table 7.2). This is physically consistent: monthly changes in water storage are governed primarily by the climatic water balance and by the persistence (memory) of the storage anomaly itself. The successful transfer of the model to the Seine basin – a markedly different hydroclimatic context from the Duero – confirms the robustness of the architecture and its potential as a transferable tool for reconstructing and forecasting groundwater and total water storage dynamics.

Table 7.2. Leading predictors of the total water storage change in the Seine by permutation importance for the selected autoregressive, late-fusion STT model. Importance is the increase in mean squared error when a predictor is randomly permuted (higher = more influential).

Type	Feature	Importance
Dynamic	Precipitation	12.61
Autoregressive	Δ TWS history (AR)	10.98
Dynamic	Air temperature	10.72
Dynamic	Population (GHS-POP)	0.60
Dynamic	CORINE land cover, 1st class	0.46
Dynamic	CORINE land cover, 2nd class	0.29
Static	Soil organic carbon (SoilGrids)	0.021
Spatial	Coordinates (X, Y)	0.016
Static	Silt content (SoilGrids)	0.015
Static	Lithology, 2nd class	0.014

8. What-if scenarios to explore future water resources

8.1. Introduction

This chapter evaluates the response of the ConvLSTM surrogate models presented in Chapters 2 and 3 under synthetic what-if forcing experiments designed to evaluate the physical consistency and generalization capacity of the trained frameworks beyond the conditions represented in their respective training domains. Rather than assessing model performance against reference simulations or observational benchmarks, the experiments presented here are oriented toward a fundamentally different question: do the surrogate models respond to imposed hydroclimatic perturbations in ways that are physically interpretable and directionally coherent with process understanding?

The what-if scenario framework relies on the application of idealized, controlled perturbations to the meteorological forcing fields that drive each model, in a way we can assess the sensitivity of predicted SWE and WTD under a range of hypothetical hydroclimatic conditions that may not be well represented in historical records. For SWE, the scenario experiments focused on the Drammen river basin and encompass gradual and transient changes in temperature, including progressive warming and cooling trends applied over the full simulation period, alongside precipitation-based perturbations ranging from sustained multi-year deficits to stepwise increases, with stochastic noise superimposed on each scenario to characterize the uncertainty envelope of the surrogate response.

For WTD, a complementary set of synthetic scenarios is applied over the Central European domain using a seasonally delta method, constructing compound wet and dry forcing states by simultaneously perturbing precipitation and soil moisture anomalies at multiples of their local climatological standard deviation, enabling a controlled decomposition of the relative contributions of precipitation forcing and antecedent soil moisture state to groundwater dynamics.

Together, these experiments serve a dual purpose within the STARS4Water modelling framework. First, they provide a structured assessment of whether the deep learning surrogates have internalized physically meaningful relationships between meteorological forcing and hydrological response, and second, to illustrate how the surrogate models can be used as computationally efficient tools for rapid scenario exploration, supporting the kind of sensitivity analysis that would be prohibitively expensive with the underlying physically based models at their native spatial resolutions.

8.2. Snow Water Equivalent

The sensitivity of the ConvLSTM surrogate model to idealized perturbations in meteorological forcing was assessed through six synthetic scenarios applied over the Drammen river basin for the period 2010–2021 (Chapter 2). The scenarios encompass three temperature-based perturbations: (i) a gradual cooling of -5°C over the full period; (ii) a gradual warming of $+5^{\circ}\text{C}$ applied over the full period; (iii) a transient warming of $+5^{\circ}\text{C}$ sustained over the first five years followed by a return to baseline conditions; and three precipitation-based perturbations: (iv) a linearly decreasing trend reaching -70% of baseline precipitation; (v) a linearly increasing trend reaching $+70\%$; (vi) and a prolonged drought imposing a sudden 70% reduction sustained over five consecutive years. In all cases, stochastic perturbations were superimposed on the forcing fields to represent uncertainty in the input data. For all scenarios, the results are evaluated against the unperturbed baseline simulation.

Figure 8.1 illustrates the domain-averaged monthly SWE time series for the six synthetic scenarios compared with the unperturbed baseline simulation. The results reveal a pronounced sensitivity of the simulated snowpack to temperature perturbations, while precipitation perturbations also produce substantial but more temporally heterogeneous responses. Under the gradual cooling (-5°C) scenario, peak SWE increases progressively throughout the simulation period, with annual maxima exceeding 500 mm toward the end of the period and reaching values above 600 mm in the final years. In contrast, gradual warming ($+5^{\circ}\text{C}$) progressively suppresses the seasonal snow cycle. The divergence from the baseline becomes increasingly pronounced over time, with SWE peaks declining to very low values during the final years of the simulation, indicating a substantial reduction in snowpack persistence under sustained warming.

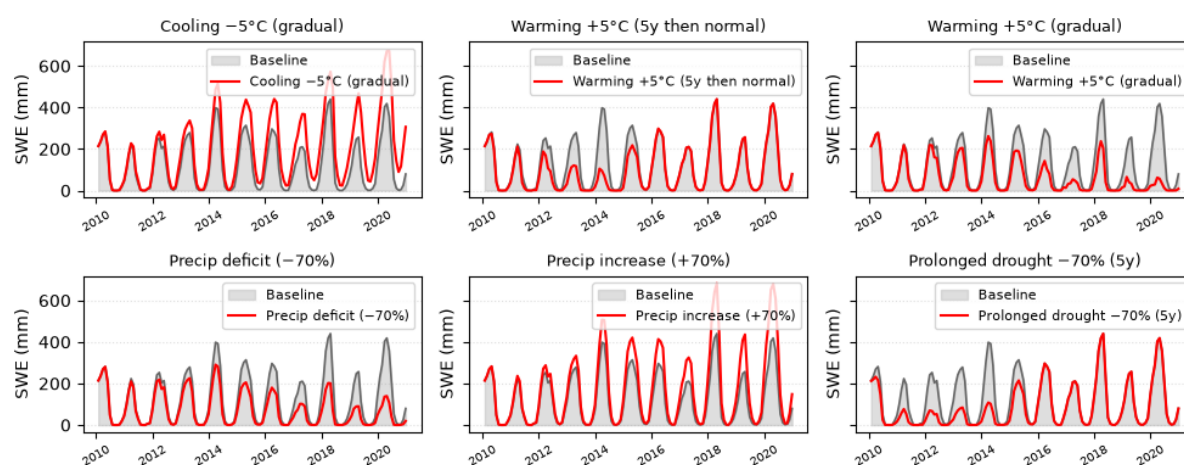


Figure 8.1. Domain-averaged monthly snow water equivalent (SWE) time series (2010–2021) for six what-if scenarios (red) compared to the baseline simulation (grey shading).

The transient warming scenario ($+5^{\circ}\text{C}$ followed by a return to baseline conditions) illustrates the reversibility of the model response to the imposed forcing perturbation. During the initial warming period, SWE peaks are substantially reduced relative to the baseline. Following the removal of the temperature perturbation, the simulated snowpack progressively recovers, with the seasonal dynamics returning toward, and in some years exceeding, the magnitudes of the unperturbed simulation. This response indicates that the surrogate model is capable of propagating both the onset and subsequent removal of a sustained temperature perturbation through the predicted snowpack dynamics. The precipitation scenarios also produce discernible changes in SWE, although their effects are less systematically amplified than those associated with the temperature perturbations. Under the -70% linear precipitation-deficit scenario, SWE peaks are generally reduced relative to the baseline, particularly during the latter part of the simulation period. Conversely, the $+70\%$ precipitation-increase scenario produces several SWE peaks that exceed the baseline, with the differences becoming particularly pronounced in the later years. The prolonged drought scenario, in which precipitation is suddenly reduced by 70% for five consecutive years, produces a marked suppression of SWE during the imposed drought period, followed by a recovery after precipitation returns to baseline conditions.

The results indicate that the ConvLSTM surrogate responds coherently to both temperature and precipitation perturbations, while the magnitude and persistence of the response depend on the type and temporal structure of the forcing perturbation. The comparatively strong response to temperature perturbations is consistent with the dominant role of temperature in controlling snow accumulation and melt processes. The precipitation response should not, however, be interpreted solely in terms of physical importance. Differences in the sensitivity of the surrogate to individual forcings may also reflect the statistical relationships learned during training and the relative explanatory power of each input variable in the training dataset.

Figure 8.2 compares the annual mean SWE anomalies of each scenario against the baseline day-of-year climatology, with the corresponding baseline anomalies shown for reference. Overall, the results show that the ConvLSTM surrogate responds consistently to the imposed perturbations, with the magnitude and temporal evolution of the anomalies depending on the type and temporal structure of the forcing. Temperature perturbations produce the most pronounced and systematic responses. Under gradual cooling (-5°C), SWE anomalies show an overall increasing tendency, exceeding $+100$ mm in several later years and reaching approximately $+180$ mm in the final year. Conversely, gradual warming ($+5^{\circ}\text{C}$) produces increasingly negative anomalies, reaching approximately -100 mm toward the end of the simulation.

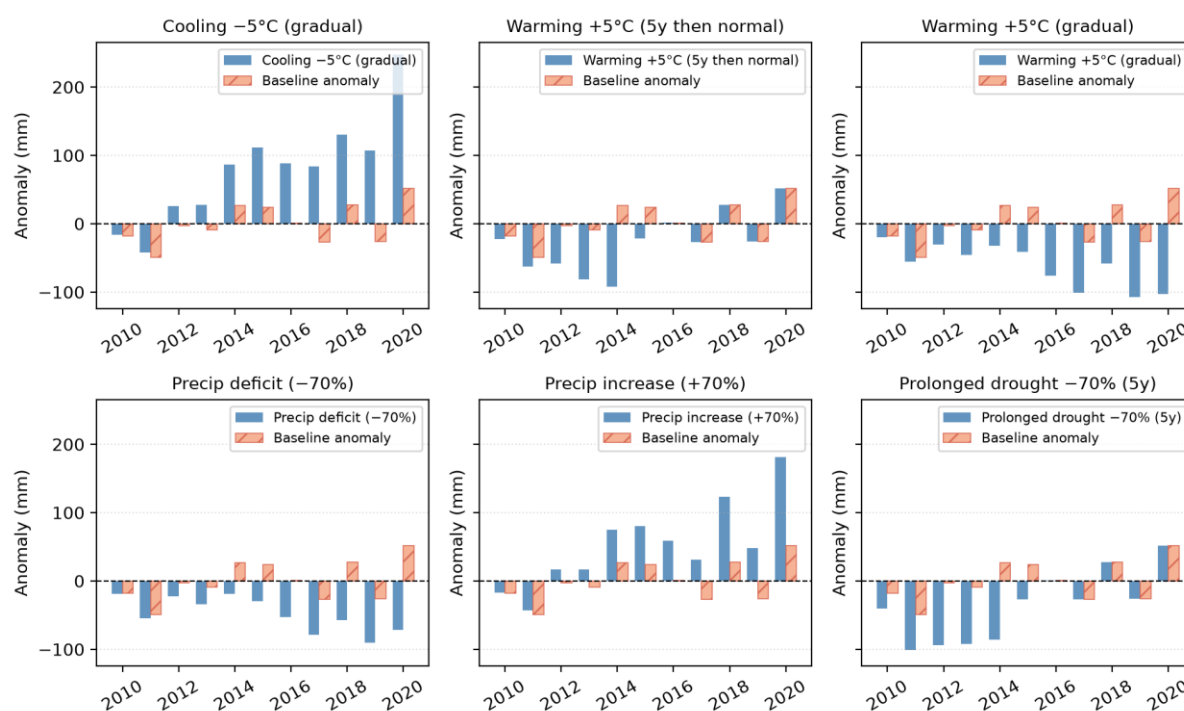


Figure 8.2. Annual anomalies of domain-averaged snow water equivalent (SWE) for each perturbation scenario relative to the baseline (2010–2021).

The transient warming scenario produces pronounced negative anomalies during the initial five-year perturbation period, followed by a reduction in anomaly magnitude after the forcing returns to baseline conditions. The precipitation experiments show more variable responses. The linear precipitation-deficit (-70%) scenario produces predominantly negative anomalies, particularly during the later years as the imposed deficit increases, whereas the precipitation-increase ($+70\%$) scenario produces predominantly positive anomalies during the latter part of the simulation, with several years

exceeding +50 mm. These results indicate that the response to precipitation perturbations depends on both their magnitude and temporal evolution rather than being negligible.

The prolonged drought scenario produces the clearest precipitation-driven response, with strongly negative SWE anomalies during the five-year period of imposed precipitation reduction, reaching approximately –100 mm in some years. After precipitation returns to baseline conditions, the anomalies progressively move toward zero and become positive in some subsequent years, indicating a recovery of simulated SWE toward the baseline regime. Overall, the figure demonstrates that the surrogate is sensitive to both the magnitude and temporal structure of the perturbations, with sustained temperature changes producing the clearest long-term response and finite-duration perturbations showing a subsequent adjustment after the forcing is removed.

Figure 8.3 shows the annual accumulation and melt components of the SWE water balance for each scenario. Across all cases, both components exhibit marked year-to-year variability driven by the natural climate variability of the baseline forcing, and none of the imposed perturbations produce a clear directional trend in the net snow balance over the simulated period. Under cooling, accumulation and melt both increase relative to baseline, reflecting a longer snow season, while the net balance remains near zero. Under gradual warming, accumulation declines as more winter precipitation falls as rain and snow melt accelerates, leading to increasingly negative net balances in later years. The transient warming scenario follows the same pattern during the perturbed period, with both components recovering once temperatures return to baseline levels. For precipitation scenarios, the perturbation signal appears almost entirely in the accumulation component, with melt fluxes staying close to baseline values, suggesting that precipitation amount primarily controls how much snow is available rather than the melt process itself. The prolonged drought scenario shows the strongest suppression of accumulation during the first five years, followed by a gradual recovery of both components once the drought period ends.

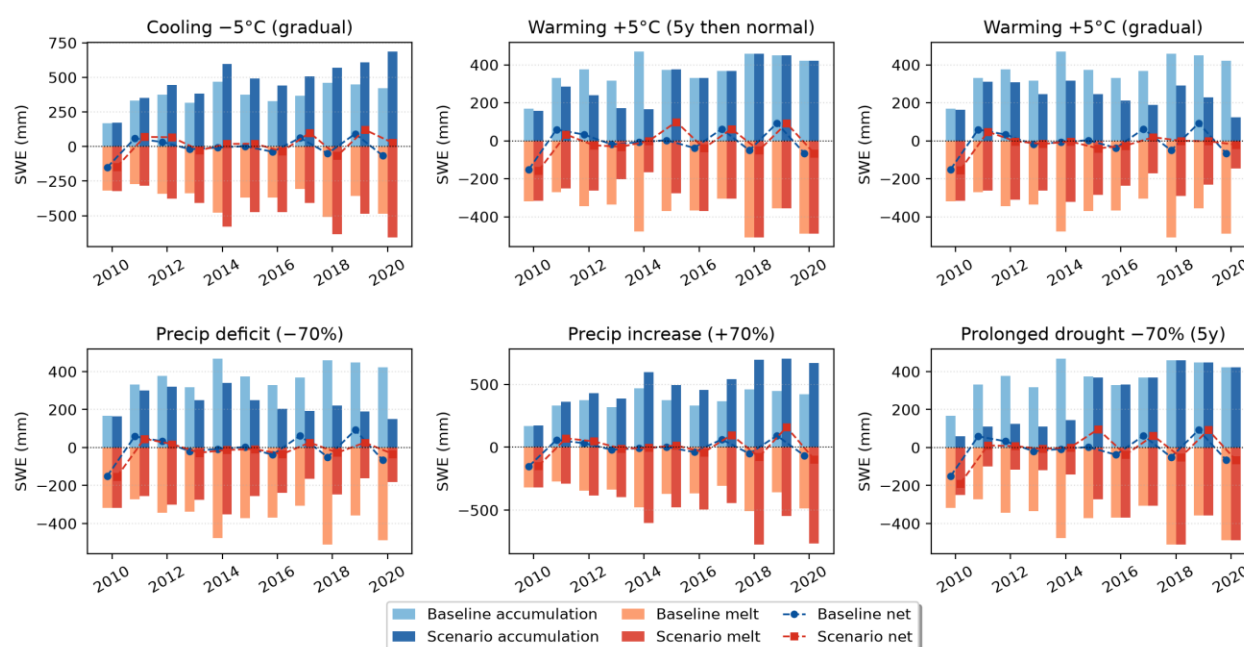


Figure 8.3. Annual SWE water balance decomposition into accumulation and melt components for baseline and each climate scenario (2010–2021).

Figure 8.4 shows the spatial distribution of the mean SWE anomaly across the Drammen basin for each scenario, averaged over the full simulation period. The cooling and gradual warming scenarios produce spatially coherent anomaly patterns: cooling yields predominantly positive anomalies across most of the domain, whereas warming generates widespread negative anomalies. The transient warming scenario also produces predominantly negative anomalies, although with greater spatial heterogeneity. Because the spatial anomalies are averaged over the full simulation period, the patterns shown for the transient scenario integrate both the perturbed period and the subsequent return to baseline forcing and therefore should not be interpreted as representing the spatial response during the perturbation period alone.

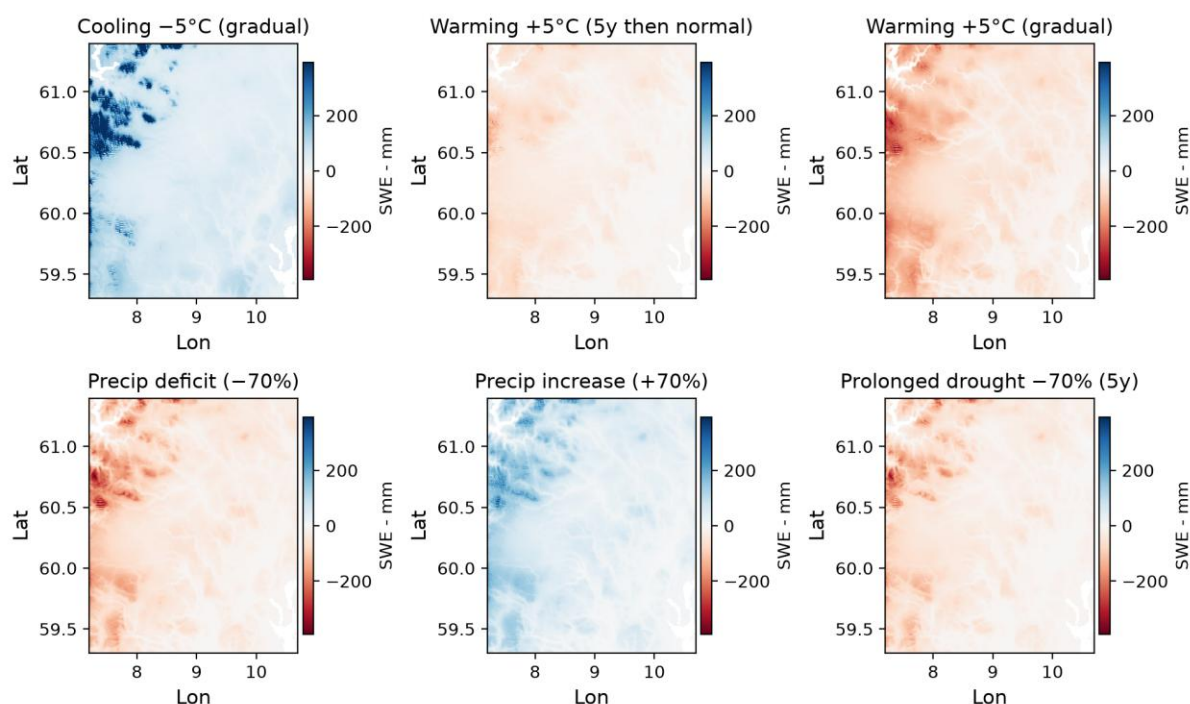


Figure 8.4. Spatial distribution of mean SWE anomaly (scenario minus baseline, in mm) across the Drammen River Basin for each climate scenario.

Among the precipitation scenarios, the +70% increase produces predominantly positive anomalies across the basin, while the -70% deficit produces mainly negative anomalies with some spatial heterogeneity. The prolonged drought scenario likewise produces predominantly negative anomalies across the domain, although the magnitude of the response varies spatially. The presence of both positive and negative anomalies within individual scenarios reflects the spatially heterogeneous response of SWE across the basin. For the five-year perturbation scenarios, the maps represent the average anomaly over the complete simulation period, including the recovery period after the perturbation is removed. Therefore, the spatial patterns should be interpreted as the multi-year mean response of the surrogate rather than as the instantaneous response to the imposed forcing.

8.3. Water Table Depth

A set of what-if experiments were conducted to assess the sensitivity of the ConvLSTM surrogate model for WTD to idealise perturbations in soil moisture and precipitation forcing over the Central European domain (Chapter 3). Four synthetic scenarios were constructed using a seasonal delta method, in which perturbations are expressed as multiples of the day-of-year (DOY) standard deviation. This approach avoids the application of climatologically inconsistent perturbations across the spatially heterogeneous domain. To distinguish between moderate and extreme conditions, two intensity levels were defined for both wet and dry regimes.

The four scenarios were defined as follows:

- **Scenario 1 (Moderate Drought):** Precipitation increased by -1σ and VWCa increased by -1σ .
- **Scenario 2 (Extreme Drought):** Precipitation increased by -2σ and VWCa increased by -2σ .
- **Scenario 3 (Moderate Flood):** Precipitation reduced by $+1\sigma$ and VWCa reduced by $+1\sigma$.
- **Scenario 4 (Extreme Flood):** Precipitation reduced by $+2\sigma$ and VWCa reduced by $+2\sigma$.

Figure 8.5 presents the seasonal-delta forcing perturbations applied to precipitation and soil moisture anomalies that define the four WTD what-if scenarios. All results are evaluated against the unperturbed baseline simulation produced by the ParFlow model, allowing a direct assessment of the surrogate's physical consistency and its ability to propagate forcing perturbations into plausible groundwater responses across diverse hydrogeological settings.

Figure 8.6 presents the monthly mean domain-averaged WTDa time series for the four synthetic scenarios compared to the unperturbed baseline over the Central European domain for the period 2012–2019. The sign convention adopted throughout is that positive WTDa values indicate a water table located deeper than the climatological mean (drier-than-normal conditions), while negative values represent a shallower-than-normal water table (wetter-than-normal conditions).

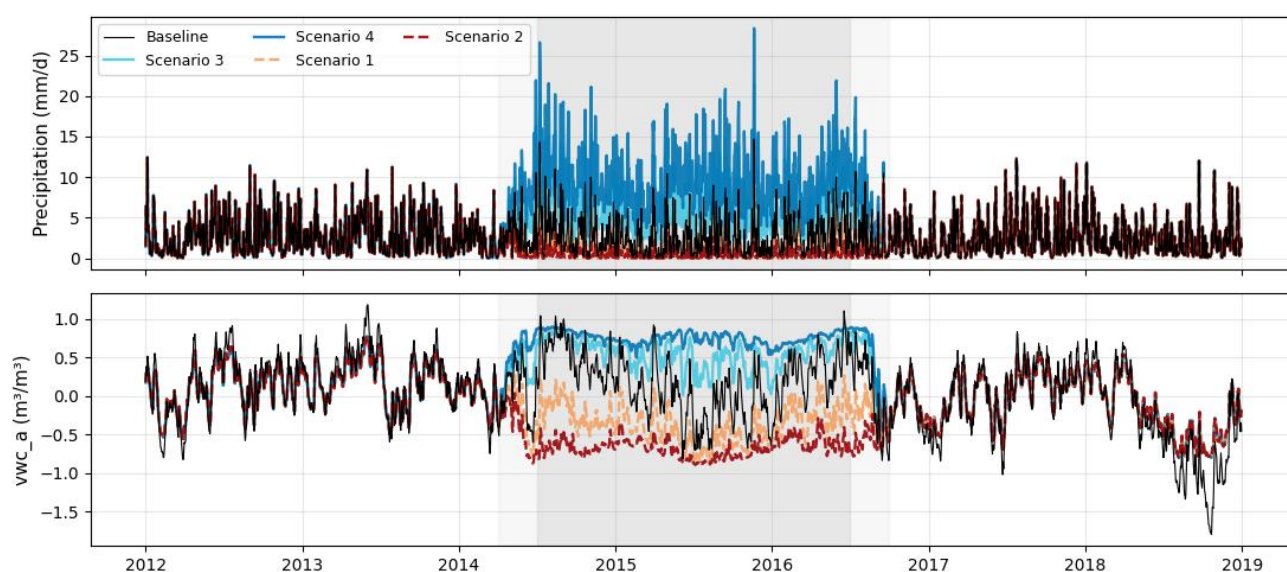


Figure 8.5. Forcing time series for the WTD what-if scenarios over Central Europe (2012–2019): daily precipitation (top) and soil moisture anomaly, VWCa (bottom), for the baseline and the four synthetic scenarios.

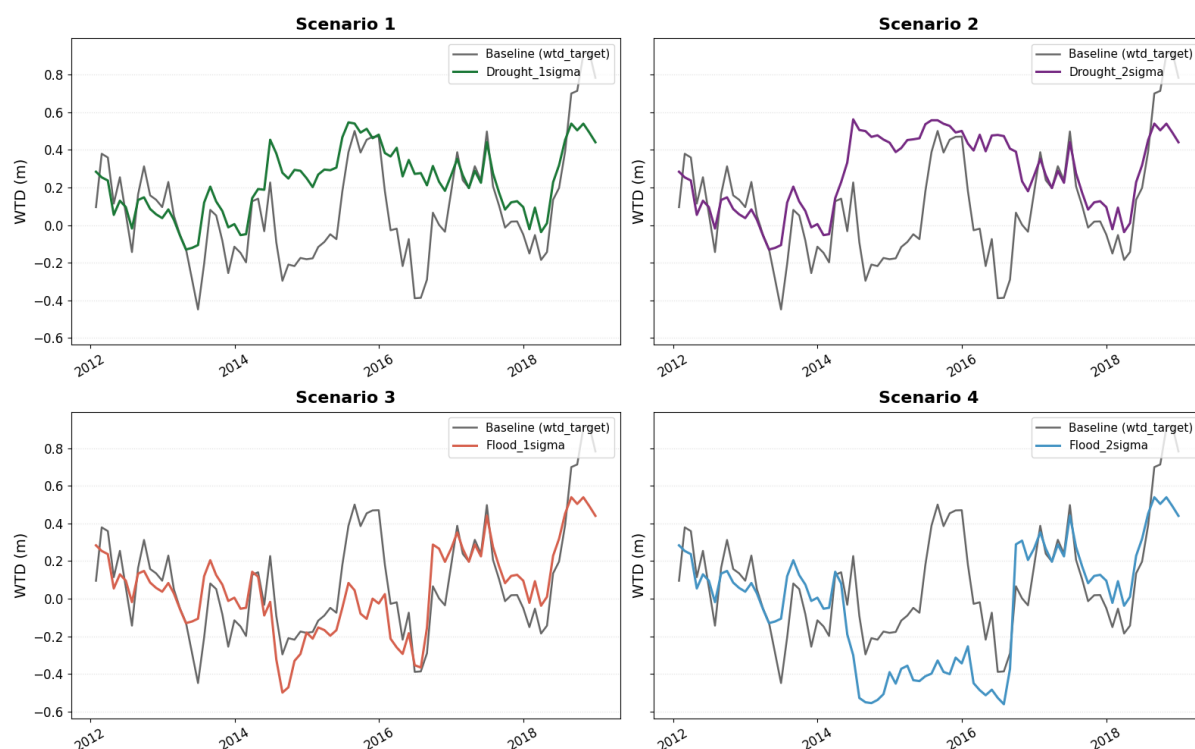


Figure 8.6. Monthly mean domain-averaged Water Table Depth (WTD) time series for the four synthetic scenarios compared against the unperturbed baseline over the Central European domain.

Scenarios 1 and 2 both produce positive WTDa departures from baseline, consistent with a deepening of the water table under dry forcing conditions. Scenario 2 generates the larger positive anomaly of the two, with domain-averaged WTDa reaching approximately +0.4 to +0.5 m during the plateau phase (2014–2016), reflecting the stronger compound signal imposed by the perturbation in both precipitation and soil moisture. Scenario 1 produces a more moderate positive anomaly, remaining predominantly in the +0.1 to +0.3 m range throughout the plateau period, consistent with the weaker +1 σ forcing applied. Both scenarios show a progressive build-up of the positive WTDa departure from 2012 onward, peak values concentrated during the plateau phase, and a gradual return toward baseline conditions by 2018–2019, in agreement with the trapezoidal structure of the imposed perturbation.

Scenarios 3 and 4 produce negative WTDa values, indicating a persistent rise of the water table toward and above the climatological mean under wet forcing conditions. Scenario 4 generates the largest negative anomaly, with domain-averaged WTDa reaching approximately –0.6 m during the plateau phase (2014–2016), driven by the simultaneous +2 σ enhancement in both precipitation and soil moisture that substantially increases recharge across the domain. Scenario 3 produces a more moderate negative departure, consistently tracking slightly above the baseline throughout the plateau period, reflecting the weaker recharge enhancement associated with the +1 σ compound wet forcing. The overall pattern across all four scenarios confirms that the ConvLSTM surrogate responds coherently to the direction and intensity of the imposed forcing perturbations, with wet scenarios producing negative WTDa and dry scenarios producing positive WTDa, and with anomaly magnitudes scaling broadly with the intensity of the applied perturbation.

Figure 8.7 presents the annual mean Water Table Depth anomaly differences (Δ WTD) between each synthetic scenario and the unperturbed baseline, aggregated by year over the 2012–2019 period. Positive Δ WTD values indicate that the scenario produces a deeper water table than the baseline (drier conditions), while negative values indicate a shallower water table (wetter conditions). The hatched orange bars represent the baseline interannual anomaly for reference.

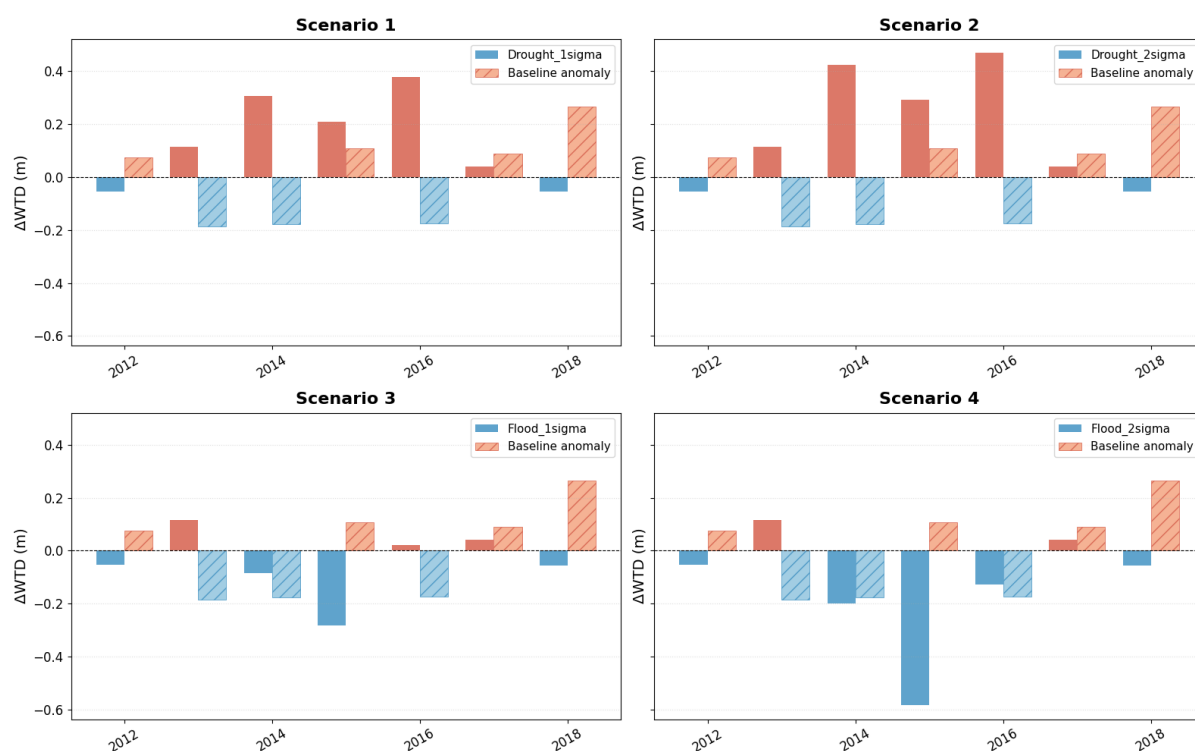


Figure 8.7. Annual WTD anomalies for each synthetic scenario relative to the baseline, with baseline interannual variability shown for comparison.

Scenarios 1 and 2 exhibit consistently positive Δ WTD departures across most years, confirming the deepening of the water table under dry forcing conditions. Scenario 1 shows modest positive departures, generally remaining below +0.2 m across all years, with the signal relatively uniform throughout the simulation period. Scenario 2 produces markedly larger positive Δ WTD values, peaking near +0.4 m in 2014 and +0.45 m in 2016, representing the strongest positive departure among the drought scenarios. Scenarios 3 and 4 produce consistently negative Δ WTD departures, with a systematic rise of the water table relative to the baseline under wet forcing. Scenario 3 reaches a peak negative departure of approximately -0.25 m around 2015, while Scenario 4 generates the most pronounced negative response, with Δ WTD reaching approximately -0.6 m in 2015, driven by the +2 σ enhancement in both precipitation and soil moisture.

Figure 8.8. presents the seasonal mean Δ WTD time series for each synthetic scenario, decomposed into the four standard meteorological seasons. Scenarios 1 and 2 produce predominantly positive Δ WTD departures across all seasons, consistent with the deepening of the water table under dry forcing conditions. In both scenarios, JJA and SON exhibit the largest positive anomalies, with peak values reaching approximately +0.4 to +0.6 m in Scenario 1 and +0.6 to +0.8 m in Scenario 2. DJF and MAM remain more moderate throughout, generally staying below +0.3 m in Scenario 1 and +0.5 m in Scenario 2. Scenario 2 consistently produces larger departures than Scenario 1 across all seasons and

years, reflecting the stronger compound drying signal of the -2σ perturbation. Both scenarios show a gradual build-up from near-zero in 2012, peak values during 2014–2016, and a convergence back toward zero by 2018–2019, consistent with the trapezoidal structure of the imposed perturbation.

Scenarios 3 and 4 produce predominantly negative Δ WTD departures, indicating a systematic shallowing of the water table under wet forcing. As in the drought scenarios, JJA and SON drive the largest anomalies, with peak negative values reaching approximately -0.3 to -0.5 m in Scenario 3 and -0.5 to -0.8 m in Scenario 4 around 2015–2016. DJF and MAM remain closer to zero or show moderate negative departures throughout most of the period. Scenario 4 consistently produces larger negative departures than Scenario 3, reflecting the stronger $+2\sigma$ compound wet forcing. The seasonal asymmetry observed across all four scenarios — with JJA and SON exhibiting systematically larger Δ WTD amplitudes than DJF and MAM — suggests that the surrogate is more sensitive to forcing perturbations during the warm season, likely reflecting the stronger coupling between soil moisture and evapotranspiration during summer and autumn that amplifies the propagation of VWC anomalies into groundwater storage changes. All seasonal signals converge toward near-zero Δ WTD by 2018–2019, consistent with the ramp-down phase of the trapezoidal perturbation.

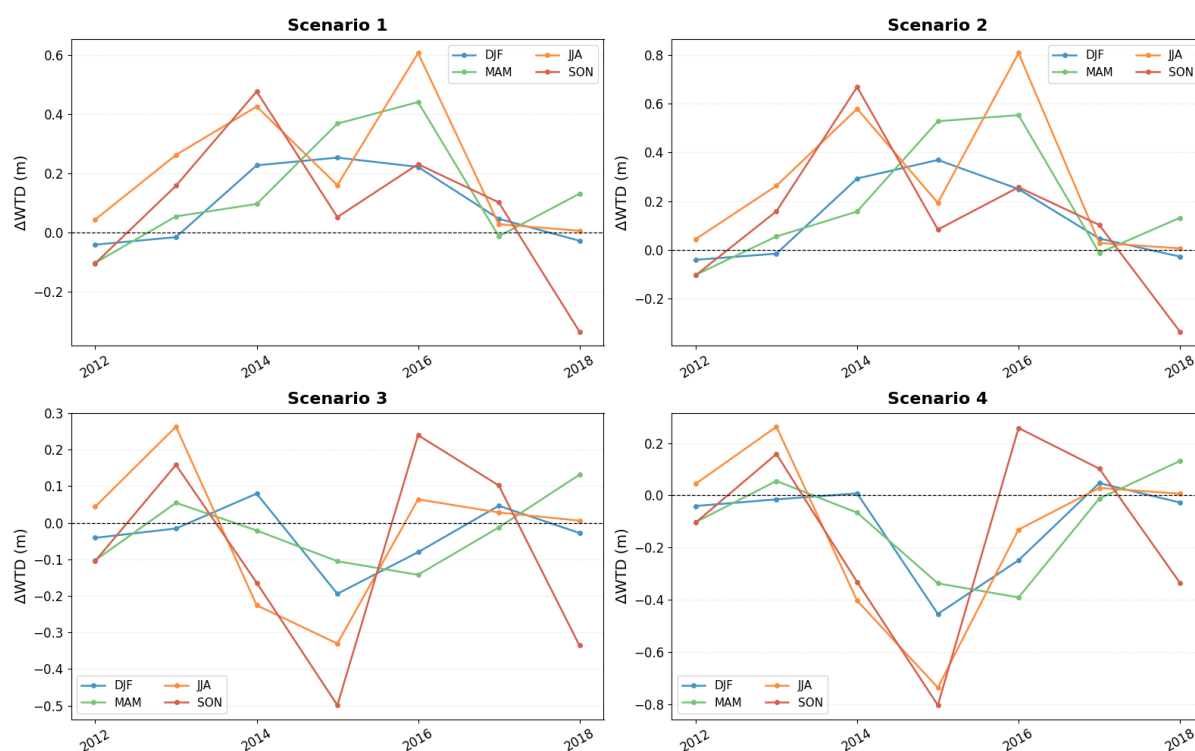


Figure 8.8. Seasonal WTD anomalies by meteorological season for each synthetic scenario, showing the intra-annual timing of groundwater response.

Finally, Figure 8.9 presents the time-mean spatial distribution of Δ WTD, averaged over the full simulation period. Scenarios 1 and 2 are both dominated by positive Δ WTD anomalies, consistent with the domain-wide drying signal discussed in Figures 8.9–8.11, but they differ clearly in the intensity and extent of the response. In both scenarios, the largest anomalies are concentrated in a dense band in the west-central portion of the domain and along the southern margin, where values exceed $+0.6$ to $+0.8$ m, while the weakest responses, including localized patches of slightly negative anomalies, appear scattered through the central and north-central portion of the domain. Scenario 2 shows a clearly stronger and more spatially coherent drying scenario, with a high-anomaly band and southern

margin are more intense, and the moderate positive anomalies (+0.2 to +0.4 m) extend across a much larger fraction of the domain, whereas in Scenario 1 these moderate anomalies are patchier and a larger area of the central-northern domain remains close to neutral or weakly negative.

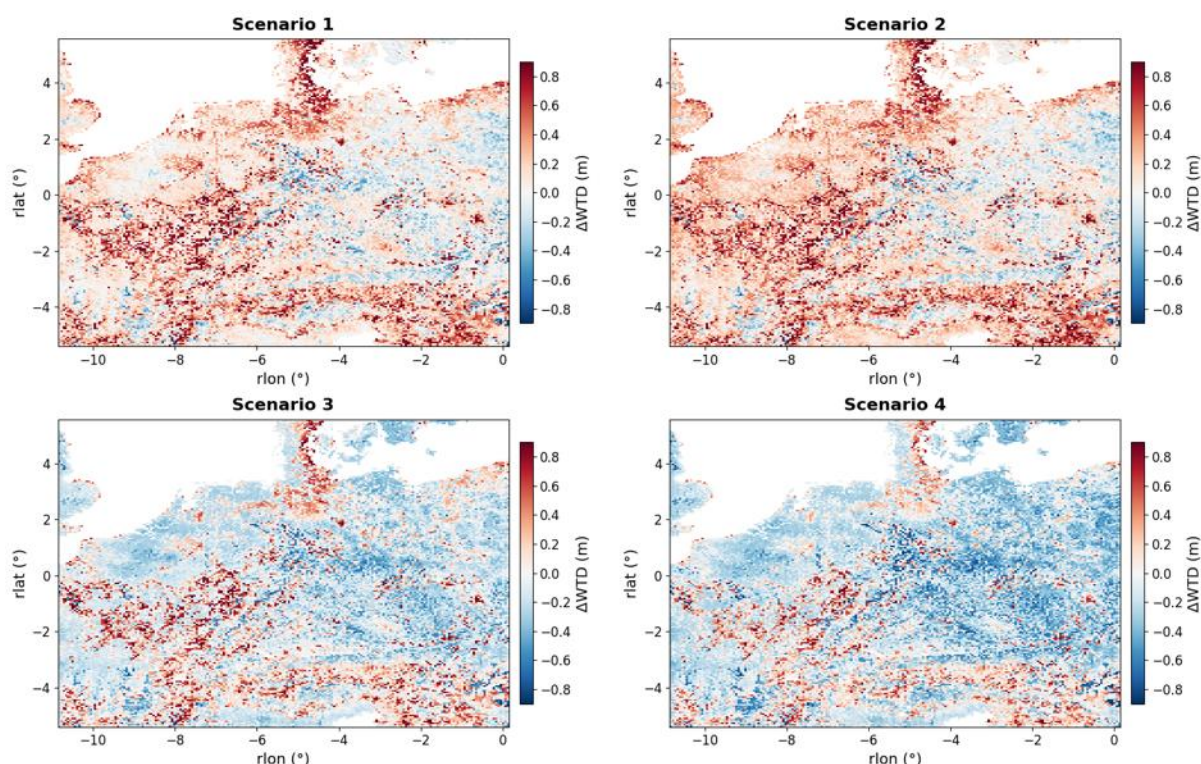


Figure 8.9. Spatial distribution of WTD anomalies across the DE06 domain for each synthetic scenario.

Scenarios 3 and 4 are both dominated by negative ΔWTD anomalies, consistent with a wetting response, but again the two differ substantially in magnitude and spatial extent. In Scenario 3, the strongest wetting signal is concentrated across the central and eastern portion of the domain, while the west-central band and southern margin retain clearly positive anomalies. In Scenario 4, the wetting signal becomes much more dominant and spatially extensive, with strong negative anomalies covering nearly the entire domain; only small, isolated patches within the west-central band and along parts of the southern margin retain weak positive anomalies.

Overall, the spatial patterns confirm that anomaly sign is controlled by the direction of the imposed perturbation, while anomaly magnitude and spatial extent scale clearly with perturbation intensity in both the dry and wet pairs. The recurring localized drying signal in the west-central band and along the southern margin, which persists to some degree across all four scenarios, points to an underlying structural feature of the domain that is comparatively robust to the magnitude of the imposed precipitation and soil moisture perturbations.

9. Conclusions and next steps

This deliverable addressed a central question for the STARS4Water project: can machine learning and deep learning models, originally developed and trained at a river basin scale, be upscaled up to national or continental domains without a substantial loss in predictive performance? Across the seven modelling components examined in this report, from deep learning surrogates for snow and groundwater dynamics to ensemble-tree reservoir models and predictive mapping tools for water quality, the results presented in this report suggest consistently toward a positive answer.

The clearest pattern emerging from this work is that models built on high-resolution physical simulations or local observational records retain much of their skill when applied to new basins, regions, or continental-scale domains. This held true regardless of the specific modelling approach: ConvLSTM surrogates for snow water equivalent and water table depth, ensemble-tree models for reservoir storage, MLMapper for groundwater nitrate contamination, and hybrid spatio-temporal architectures for groundwater storage all showed transferable behaviour. The amount of additional retraining or fine-tuning needed in each case tracked closely with how different the target domain was from the original training environment, climatologically and physiographically.

A second key finding is that globally available reanalysis forcing products, and ERA5-Land in particular, can sustain meaningful surrogate model performance even in the absence of high-resolution regional datasets. This is significant because it implies that the proposed frameworks can, in principle, be deployed across Europe and beyond without depending on the availability of dense observational networks or costly regional atmospheric simulations. At the same time, the results consistently identified targeted bias correction and fine-tuning with local data as the most impactful steps for improving skill in operationally relevant configurations, pointing to a clear and actionable pathway for further development.

The what-if scenario analyses provided an additional line of evidence for the physical consistency of the surrogate models, confirming that they respond to idealized hydroclimatic perturbations in interpretable and physically plausible ways. Temperature emerged as the dominant control on snowpack dynamics in the Drammen Basin, while for groundwater, the volumetric soil moisture perturbations produced the most spatially extensive and dominant responses across the Central European domain. These results support the use of the proposed frameworks not only for monitoring and nowcasting applications, but also for scenario-based assessments of climate change impacts on water resources.

Collectively, the results of this deliverable establish data-driven and deep learning surrogate approaches as flexible, transferable, and physically grounded components of the STARS4Water modelling framework, capable of bridging the gap between basin-scale process understanding and EU-level water resources assessment.

Looking ahead, the priority areas for future work include the application of bias correction to ERA5-Land forcing fields to recover predictive skill in regions currently limited by forcing data quality, targeted fine-tuning of the continental-scale surrogates using local in-situ observations in climatically distinct regions such as the Mediterranean and the Iberian Peninsula, the incorporation of snowmelt-related predictors to improve groundwater recharge representation in snow-influenced catchments, and the development of ensemble-based uncertainty quantification frameworks to support operational confidence assessment. More broadly, Deliverable 5.3 will focus on extending the transfer learning framework beyond Europe, exploring the applicability of the trained surrogate models in the US, assessing their potential contribution to global-scale water resources monitoring.

10. Bibliography

- Aasen, M., Westskog, H., Wilhite, H., & Lindberg, M. (2010). The EU electricity disclosure from the business perspective—A study from Norway. *Energy Policy*, 38(12), 7921–7928.
- Avila, L. & S. Kollet (2026a). Improving existing models on multiscale model integration. Deliverable D3.3. (D3.3). STARS4Water project.
- Avila, L. & S. Kollet (2026b). EU-workshop on the use of STARS4Water data-driven models. Deliverable D6.8. (D6.8). STARS4Water project.
- Baron, H. & D.J. Keller (2026). Data-driven modelling tools for the STARS4Water river basins. Deliverable D3.4. (D3.4.). STARS4Water project.
- Bair, E. H., Abreu Calfa, A., Rittger, K., & Dozier, J. (2018). Using machine learning for real-time estimates of snow water equivalent in the watersheds of Afghanistan. *The Cryosphere*, 12, 1579–1594. <https://doi.org/10.5194/tc-12-1579-2018>
- 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, 303–309. <https://doi.org/10.1038/nature04141>
- Barzegar, R., Aalami, M. T., & Adamowski, J. (2020). Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. *Stochastic Environmental Research and Risk Assessment*, 34, 415–433. <https://doi.org/10.1007/s00477-019-01776-7>
- Belleflamme, A., Goergen, K., Wagner, N., Kollet, S., Bathiany, S., El Zohbi, J., Rechid, D., Vanderborght, J., & Vereecken, H. (2023). Hydrological forecasting at impact scale: the integrated ParFlow hydrological model at 0.6 km for climate resilient water resource management over Germany. *Frontiers in Water*, 5, 1183642.
- Boergens, E., Güntner, A., Dobslaw, H., & Dahle, C. (2020). Quantifying the central European droughts in 2018 and 2019 with GRACE Follow-On. *Geophysical Research Letters*, 47, e2020GL087285.
- Brás, T., Seixas, J., Carvalhais, N., & Jägermeyr, J. (2021). Severity of drought and heatwave crop losses tripled over the last five decades in Europe. *Environmental Research Letters*, 16, 065012.
- Breiman, L. (1999). Pasting small votes for classification in large databases and on-line. *Machine Learning*, 36, 85–103.
- Broxton, P., et al. (2019). Improving snow water equivalent maps with machine learning of snow survey and lidar measurements. *Water Resources Research*, 55, 1544–1561. <https://doi.org/10.1029/2018WR024146>
- Buizza, R. (1997). Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Monthly Weather Review*, 125(1), 99–119.
- Buizza, R. (2006). The ECMWF ensemble prediction system. In *Predictability of Weather and Climate* (pp. 459–488).

- Chapman, W. E., Delle Monache, L., Alessandrini, S., Subramanian, A. C., Ralph, F. M., Xie, S. P., & Hayatbini, N. (2022). Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Monthly Weather Review*, 150(1), 215–234.
- Čistý, M., Danko, M., Kohnová, S., Považanová, B., & Trizna, A. (2024). Machine learning enhanced by feature engineering for estimating snow water equivalent. *Water*, 16, 2285. <https://doi.org/10.3390/w16162285>
- Clark, M. P., Hendriks, J., Slater, A. G., Kavetski, D., Anderson, B., Cullen, N. J., et al. (2011). Representing spatial variability of snow water equivalent in hydrologic and land-surface models: A review. *Water Resources Research*, 47, W07539. <https://doi.org/10.1029/2011WR010745>
- Dasgupta, A., Arnal, L., Emerton, R., Harrigan, S., Matthews, G., Muhammad, A., & Pappenberger, F. (2023). Connecting hydrological modelling and forecasting from global to local scales: perspectives from an international joint virtual workshop. *Journal of Flood Risk Management*, in press.
- Di Taranto, L., Fiorentino, A., Doglioni, A., & Simeone, V. (2025). Modeling water table response in Apulia (southern Italy) with global and local LSTM-based groundwater forecasting. *Water*, 17(15), 2268.
- Do, H. X., et al. (2017). A global-scale investigation of trends in annual maximum streamflow. *Journal of Hydrology*, 552, 28–43.
- Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., & Drusch, M. (2017). ESA CCI Soil Moisture for improved Earth system understanding: State-of-the-art and future directions. *Remote Sensing of Environment*, 203, 185–215.
- Dozier, J., & Painter, T. H. (2004). Multiscale imaging of snow properties using AVIRIS and MODIS. *Proceedings of the 13th AVIRIS Workshop*, 1–12.
- Egli, L., Jonas, T., & Meister, R. (2009). Comparison of different automatic methods for estimating snow water equivalent. *Cold Regions Science and Technology*, 57, 107–115. <https://doi.org/10.1016/j.coldregions.2009.01.007>
- Fassnacht, S. R., Dressler, K. A., & Bales, R. C. (2003). Snow water equivalent interpolation for the Colorado River Basin from snow telemetry (SNOTEL) data. *Water Resources Research*, 39. <https://doi.org/10.1029/2002WR001512>
- Fekete, B. M., Vörösmarty, C. J., & Grabs, W. (2007). The current status of global river discharge monitoring and potential new technologies complementing traditional discharge measurements. *Hydrological Sciences Journal*, 52(4), 633–651.
- Fontrodon-Bach, A., Schaeffli, B., Woods, R., Teuling, A. J., & Larsen, J. R. (2023). NH-SWE: Northern Hemisphere Snow Water Equivalent dataset based on in situ snow depth time series. *Earth System Science Data*, 15(6), 2577–2599.
- Furusho-Percot, C., Goergen, K., Hartick, C., Kulkarni, K., Keune, J., & Kollet, S. (2019). Pan-European groundwater to atmosphere terrestrial systems climatology from a physically consistent simulation. *Scientific Data*, 6, 320.
- Giorgi, F., et al. (2009). Addressing climate information needs at the regional level: The CORDEX framework. *WMO Bulletin*, 58(3), 175–183.

- Graabak, I., Jaehnert, S., Korpås, M., & Mo, B. (2017). Norway as a battery for the future European power system — impacts on the hydropower system. *Energies*, 10, 2054. <https://doi.org/10.3390/en10122054>
- Gutowski, W. J., et al. (2016). WCRP coordinated regional downscaling experiment (CORDEX): A diagnostic MIP for CMIP6. *Geoscientific Model Development*, 9, 4087–4095.
- Hailemariam, S. F., & Alfredsen, K. (2023). Quantitative flood risk assessment in Drammenselva River, Norway. *Water*, 15(5), 920.
- Hartick, C., Furusho-Percot, C., Goergen, K., & Kollet, S. (2021). An interannual probabilistic assessment of subsurface water storage over Europe using a fully coupled terrestrial model. *Water Resources Research*, 57, e2020WR027828.
- Hokkanen, J., Kollet, S., Kraus, J., Herten, A., Hrywniak, M., & Pleiter, D. (2021). Leveraging HPC accelerator architectures with modern techniques – hydrologic modeling on GPUs with ParFlow. *Computational Geosciences*, 25, 1579–1590.
- Hu, W., Cervone, G., Young, G., & Delle Monache, L. (2023). Machine learning weather analogs for near-surface variables. *Boundary-Layer Meteorology*, 186(3), 711.
- Jenicek, M., & Ledvinka, O. (2020). Importance of snowmelt contribution to seasonal runoff and summer low flows in Czechia. *Hydrology and Earth System Sciences*, 24, 3475–3491. <https://doi.org/10.5194/hess-24-3475-2020>
- Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremier, D., Weisheimer, A., Balsamo, G., et al. (2019). SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, 12(3), 1087–1117.
- Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts. *Journal of Hydrology*, 320, 173–186. <https://doi.org/10.1016/j.jhydrol.2005.07.013>
- Khorram, S., & Jehbez, N. (2023). A hybrid CNN-LSTM approach for monthly reservoir inflow forecasting. *Water Resources Management*, 37(10), 4097–4121.
- Kim, K. B., Kwon, H. H., & Han, D. (2022). Intercomparison of joint bias correction methods for precipitation and flow from a hydrological perspective. *Journal of Hydrology*, 604, 127261.
- King, F., Erler, A. R., Frey, S. K., & Fletcher, C. G. (2020). Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada. *Hydrology and Earth System Sciences*, 24, 4887–4902. <https://doi.org/10.5194/hess-24-4887-2020>
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Kjaerland, F. (2007). A real option analysis of investments in hydropower — the case of Norway. *Energy Policy*, 35, 5901–5908. <https://doi.org/10.1016/j.enpol.2007.06.019>
- Knoben, W. J. (2024). Setting expectations for hydrologic model performance with an ensemble of simple benchmarks. *Hydrological Processes*, 38(10), e15288.
- 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, 945–958.

- Kollet, S. J., & Maxwell, R. M. (2008). Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model. *Water Resources Research*, 44(2).
- Kollet, S. J., Gasper, F., Brdar, S., Goergen, K., Hendricks-Franssen, H.-J., Keune, J., & Shrestha, P. (2018). Introduction of an experimental terrestrial forecasting/monitoring system at regional to continental scales based on the Terrestrial Systems Modeling Platform (v1.1.0). *Water*, 10, 1697.
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22, 6005–6022. <https://doi.org/10.5194/hess-22-6005-2018>
- 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, 5089–5110.
- Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. *Hydrology and Earth System Sciences*, 28(17), 4187–4201.
- Kuffour, B. N. O., Engdahl, N. B., Woodward, C. S., Condon, L. E., Kollet, S., & Maxwell, R. M. (2020). Simulating coupled surface-subsurface flows with ParFlow v3.5.0: capabilities, applications, and ongoing development of an open-source, massively parallel, integrated hydrologic model. *Geoscientific Model Development*, 13, 1373–1397.
- Le, T., & Behrangi, A. (2022). Deep learning for precipitation estimation: A review and new perspectives. *Environmental Modelling & Software*, 148, 105239. <https://doi.org/10.1016/j.envsoft.2021.105239>
- Leisenring, M., & Moradkhani, H. (2011). Snow water equivalent prediction using Bayesian data assimilation methods. *Stochastic Environmental Research and Risk Assessment*, 25, 253–270. <https://doi.org/10.1007/s00477-010-0445-5>
- Li, X., et al. (2021). Improved daily SMAP satellite soil moisture prediction over China using a deep learning model with transfer learning. *Remote Sensing*, 13(4), 615.
- Liang, Z., Liu, Y., Hu, H., Li, H., Ma, Y., & Khan, M. Y. A. (2021). Combined wavelet transform with long short-term memory neural network for water table depth prediction in Baoding City, North China Plain. *Frontiers in Environmental Science*, 9, 780434.
- Lussana, C., Saloranta, T., Skaugen, T., Magnusson, J., Tveito, O. E., & Andersen, J. (2018). seNorge2 daily precipitation, an observational gridded dataset over Norway from 1957 to the present day. *Earth System Science Data*, 10(1), 235–249.
- Lussana, C., Tveito, O. E., Dobler, A., & Tunheim, K. (2019). seNorge_2018, daily precipitation, and temperature datasets over Norway. *Earth System Science Data*, 11(4), 1531–1551.
- Ma, K., Feng, D., Lawson, K., Tsai, W. P., Liang, C., Huang, X., & Shen, C. (2021). Transferring hydrologic data across continents – Leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, 57(5), e2020WR028600.
- Ma, Y., Montzka, C., Naz, B., & Kollet, S. (2022). Advancing AI-based pan-European groundwater monitoring. *Environmental Research Letters*, 17, 114037.

- Ma, Y., Leonarduzzi, E., Defnet, A., Melchior, P., Condon, L. E., & Maxwell, R. M. (2024). Water table depth estimates over the contiguous United States using a random forest model. *Groundwater*, 62(1), 34–43.
- Margulis, S. A., Giroto, M., Cortés, G., & Durand, M. (2019). A particle batch smoother approach to snow water equivalent estimation. *Journal of Hydrometeorology*, 20, 937–955. <https://doi.org/10.1175/JHM-D-18-0189.1>
- 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, 923–937.
- McCuen, R. H., Knight, Z., & Cutter, A. G. (2006). Evaluation of the Nash–Sutcliffe efficiency index. *Journal of Hydrologic Engineering*, 11(6), 597–602.
- Mital, U., et al. (2022). Modeling spatial distribution of snow water equivalent by combining meteorological and satellite data with lidar maps. *Artificial Intelligence for the Earth Systems*, 1, e2022-AIES-D-22-0010.1. <https://doi.org/10.1175/AIES-D-22-0010.1>
- Moishin, M., Deo, R. C., Prasad, R., Raj, N., & Abdulla, S. (2021). Designing deep-based learning flood forecast model with ConvLSTM hybrid algorithm. *IEEE Access*, 9, 50982–50993.
- Molotch, N. P., & Bales, R. C. (2005). Scaling snow observations from the point to the grid element: Implications for observation network design. *Water Resources Research*, 41, W11421. <https://doi.org/10.1029/2005WR004229>
- Mooney, P. A., & Li, L. (2021). Near future changes to rain-on-snow events in Norway. *Environmental Research Letters*, 16, 064039. <https://doi.org/10.1088/1748-9326/abf93f>
- Moravec, V., Markonis, Y., Rakovec, O., Svoboda, M., Trnka, M., Kumar, R., & Hanel, M. (2021). Europe under multi-year droughts: how severe was the 2014–2018 drought period? *Environmental Research Letters*, 16, 034062.
- Muñoz-Sabater, J., Lawrence, H., Albergel, C., de Rosnay, P., Isaksen, L., Mecklenburg, S., Kerr, Y., & Drusch, M. (2019). Assimilation of SMOS brightness temperatures in the ECMWF IFS. *ECMWF Technical Memorandum*, 843, 38 pp.
- 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, 418–424. <https://doi.org/10.1038/s41558-021-01014-9>
- Naumann, G., Cammalleri, C., Mentaschi, L., & Feyen, L. (2021). Increased economic drought impacts in Europe with anthropogenic warming. *Nature Climate Change*, 11, 485–491.
- Ntokas, K. F. F., Odry, J., Boucher, M. A., & Garnaud, C. (2021). Investigating ANN architectures and training to estimate snow water equivalent from snow depth. *Hydrology and Earth System Sciences*, 25, 3017–3040. <https://doi.org/10.5194/hess-25-3017-2021>
- O'Neill, M., Tijerina, D., Condon, L., & Maxwell, R. (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, 7223–7254.
- Pall, P., Tallaksen, L. M., & Stordal, F. (2019). A climatology of rain-on-snow events for Norway. *Journal of Climate*, 32, 6995–7016. <https://doi.org/10.1175/JCLI-D-18-0529.1>

- Patakchi Yousefi, K., & Kollet, S. (2023). Deep learning of model- and reanalysis-based precipitation and pressure mismatches over Europe. *Frontiers in Water*, 5, 1178114. <https://doi.org/10.3389/frwa.2023.1178114>
- Ryd, E., & Nearing, G. (2025). Fine flood forecasts: Incorporating local data into global models through fine-tuning. *arXiv preprint arXiv:2504.12559*.
- Saloranta, T. M. (2012). Simulating snow maps for Norway: description and statistical evaluation of the seNorge snow model. *The Cryosphere*, 6, 1323–1337. <https://doi.org/10.5194/tc-6-1323-2012>
- Saloranta, T. M. (2016). Operational snow mapping with simplified data assimilation using the seNorge snow model. *Journal of Hydrology*, 538, 314–325.
- Shao, D., Li, H., Wang, J., Hao, X., Che, T., & Ji, W. (2022). Reconstruction of a daily gridded snow water equivalent product for the land region above 45°N based on a ridge regression machine learning approach. *Earth System Science Data*, 14, 795–809. <https://doi.org/10.5194/essd-14-795-2022>
- Sheffield, J., Wood, E. F., Pan, M., Beck, H., Coccia, G., Serrat-Capdevila, A., & Verbist, K. (2018). Satellite remote sensing for water resources management: Potential for supporting sustainable development in data-poor regions. *Water Resources Research*, 54, 9724–9758. <https://doi.org/10.1029/2017WR022437>
- Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28.
- Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., & Demir, I. (2020). A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology*, 82, 2635–2670. <https://doi.org/10.2166/wst.2020.369>
- Skaugen, T., Luijting, H., Saloranta, T., Vikhamar-Schuler, D., & Müller, K. (2018). In search of operational snow model structures for the future—comparing four snow models for 17 catchments in Norway. *Hydrology Research*, 49(6), 1929–1945.
- Solvang, E., Charmasson, J., Sauterleute, J., Harby, A., Killingtveit, Å., Egeland, H., Andersen, O., Ruud, A., & Aas, Ø. (2014). *Norwegian hydropower for large-scale electricity balancing needs*. SINTEF Energy Research, Trondheim, Norway.
- Sun, R., Pan, B., & Duan, Q. (2023). A surrogate modeling method for distributed land surface hydrological models based on deep learning. *Journal of Hydrology*, 624, 129944. <https://doi.org/10.1016/j.jhydrol.2023.129944>
- Taheri, M., & Mohammadian, A. (2022). An overview of snow water equivalent: Methods, challenges, and future outlook. *Sustainability*, 14, 11395. <https://doi.org/10.3390/su141811395>
- Teutschbein, C., & Seibert, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, 456, 12–29.
- Vu, M. T., Jardani, A., Massei, N., & Fournier, M. (2021). Reconstruction of missing groundwater level data by using Long Short-Term Memory (LSTM) deep neural network. *Journal of Hydrology*, 597, 125776.

- Wang, W., Yu, Z., Zhang, W., Shao, Q., Zhang, Y., Luo, Y., & Xu, J. (2014). Responses of rice yield, irrigation water requirement and water use efficiency to climate change in China: Historical simulation and future projections. *Agricultural Water Management*, 146, 249–261.
- Wang, Y., Feng, D., & Liu, J. (2022). LSTM-based streamflow prediction: A review from hydrological perspective. *Water*, 14, 54. <https://doi.org/10.3390/w14010054>
- Winkler, M., Schellander, H., & Gruber, S. (2021). Snow water equivalents exclusively from snow depths and their temporal changes: the Δ snow model. *Hydrology and Earth System Sciences*, 25(3), 1165–1187.
- Wu, H., Yang, Q., Liu, J., & Wang, G. (2020). A spatiotemporal deep fusion model for merging satellite and gauge precipitation in China. *Journal of Hydrology*, 584, 124664. <https://doi.org/10.1016/j.jhydrol.2020.124664>
- Xie, W., Yi, S., Leng, C., Xia, D., Li, M., Zhong, Z., & Ye, J. (2022). The evaluation of IMERG and ERA5-Land daily precipitation over China with considering the influence of gauge data bias. *Scientific Reports*, 12(1), 8085.
- Xu, Y., Lin, K., Hu, C., Wang, S., Wu, Q., Zhang, L., & Ran, G. (2023). Deep transfer learning based on transformer for flood forecasting in data-sparse basins. *Journal of Hydrology*, 625, 129956.
- Yao, S., He, Y., Zhang, L., Yang, W., Chen, Y., Sun, Q., Zhao, Z., & Cao, S. (2023). A convLSTM neural network model for spatiotemporal prediction of mining area surface deformation based on SBAS-InSAR monitoring data. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–22.
- Zhang, Y., et al. (2018). Developing a long short-term memory (LSTM)-based model for predicting water table depth in agricultural areas. *Journal of Hydrology*, 561, 918–929.