STARS4Water

Data-driven modelling tools for the STARS4Water river basins.

Deliverable 3.4

September 2025

Data-driven modelling tools for STARS4Water river basins

Lead beneficiary	UK Centre for Ecology & Hydrology
Lead author(s)	Helen Baron, Virginie Keller
Work Package	3: Developing next generation models
Due date	September 2025
Submission date	September 2025
Contributors	Daniel Klotz; Annine Kenne; Leandro Avila; Devi Purnamasari; Pedro Martínez Santos; Héctor Aguilera; Víctor Gómez Escalonilla; Manuel Rodríguez del Rosario; Silvia Díaz Alcaide; Joost Beckers; Trine Jahr Hegdahl

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	23/11/2023	Helen Baron	Outline
0.2	05/01/2024	Helen Baron	Tool scoping
0.3	28/03/2024	Helen Baron	Milestone – proof of concept
0.4	31/03/2025	Helen Baron	Milestone - model progression
0.5	25/07/2025	Helen Baron	Initial review from Trine Jahr Hegdahl & Virginie Keller
0.6	15/08/2025	Helen Baron	Edits from first review
0.7	25/08/2025	Helen Baron	Second review from Trine Jahr Hegdahl
0.8	19/09/2025	Helen Baron	Final edits and formatting
1.0	29/09/2025	Harm Duel	Approval

Citation:

Baron, H., V. Keller, D. Klotz, A. Kenne, L. Avila, D, Purnamasari, P. Martínez Santos, H. Aguilera, V. Gómez Escalonilla, M. Rodríguez del Rosario, S. Díaz Alcaide, J. Beckers and T. Jahr Hegdahl, 2025. Data-driven modelling tools for STARS4Water river basins. Horizon Europe project STARS4Water. Deliverable D3.4.

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 describes the data-driven modelling tools which have been developed to help address the needs of the 7 STARS4Water river basin hubs, as detailed in report D1.2. In this work, data-driven methods have been used to leverage the increasing volumes of hydrological, environmental, and socio-economic data available to provide valuable information on available water resources and related topics. These methods are well suited to solve these complex problems, since they are able to capture non-linear patterns while assuming no prior knowledge, and can do so in a computationally efficient manner. Thus, these tools perfectly complement the development of existing processed-based modelling tools and frameworks detailed in report D3.2.

These data-driven tools were proposed and selected in a series of workshops with the project team, based on the requirements of the stakeholders and modelling gaps identified in report D3.1, and consist of:

- Prediction of reservoir storage and inflows;
- Total water storage downscaling;
- Agricultural water use;
- Predictive mapping of groundwater quality;
- Quantitative groundwater resources estimation.

These tools cover a range of water resource related issues including: the estimation and forecasting of key water sources such as reservoir and groundwater stores at a range of spatial and temporal scales; spatial prediction of groundwater contamination; and estimation of irrigated area.

This report details the modelling tools developed in this work package and analyses their application over selected catchments within the river basin hubs. Promising results are demonstrated for these tools in the selected basins, and with continued collaboration with the basin stakeholders have the potential to be deployed operationally to aid water management decisions.

Each tool also has the capacity to be extended to other regions; this will be further explored in future work within the STARS4Water project (Task 4.4).

Table of Contents

List of Acronyms

ABC	AdaBoost Classifier
AUC	Area Under the Receiver Operating Characteristic Curve
CLM	Community Land Model
CRP(S)S	Continuous Ranked Probability (Skill) Score
DL	Deep Learning
EO	Earth Observation
ET(C)	Extra Trees (Classifier)
GBC	Gradient Boosting Classifier
GRACE	Gravity Recovery and Climate Experiment
GWL	Ground Water Level
GWSC	Groundwater Storage Change
KGE	Kling-Gupta Efficiency
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Square Error
MTL	Multi-task learning
NDVI	Normalized Difference Vegetation Index
NSE	Nash-Sutcliffe Efficiency
RBH	River Basin Hub
RF(C)	Random Forest (Classifier)
RMSE	Root Mean Square Error
STARS4Water	Supporting Stakeholders for Adaptive, Resilient and Sustainable Water management
STT	Spatiotemporal Transformer
TSMP	Terrestrial Systems Modelling Platform
TWS(A)	Total Water Storage (Anomaly)
WTD(A)	Water Table Depth (Anomaly)

1 Introduction

1.1 STARS4Water project

STARS4Water (Supporting Stakeholders for Adaptive, Resilient and Sustainable Water management) is a 4 year research project under the Horizon Europe Program. The project is addressing a call on improved understanding, observation and monitoring of water resources availability to support the European Green Deal and EU water policies. The project aims to:

- improve the understanding of climate change impacts on water resources availability and the vulnerabilities for ecosystems, society, and economic sectors at river basin scale
- develop and deliver new data services and data driven models for better supporting the decision making and planning on actions for adaptive, resilient and sustainable management of freshwater resources.

The STARS4Water project includes two distinctive elements, depicted in Figure 1. Firstly, the project team works with 7 river basin hubs (RBHs) through a co-creation, living lab-type approach. The location of the selected RBHs is presented in Figure 2. The new services and models are co-designed with stakeholders to meet their needs on data and information, ensuring relevance and uptake beyond the lifetime of the project. Secondly, the team advances the use of new datasets and models and integrates these into current river basin management information tools and decision-making processes. New datasets and models offer possibilities for improved projections on water resources availability, and new insights on the links between water, nature and society allow for a broader set of indicators to inform decision-making on water management. The consortium strongly believe that this combination of stakeholder-driven and science-driven approaches will yield significant progress in climate change adaptation with respect to water resources management.

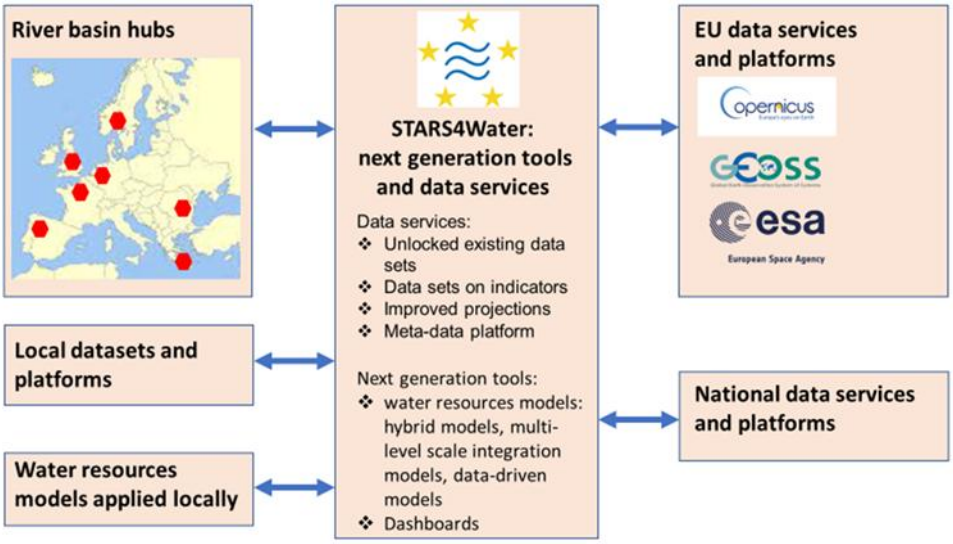


Figure 1. Overview of STARS4Water activities within the context of stakeholders and data providers.

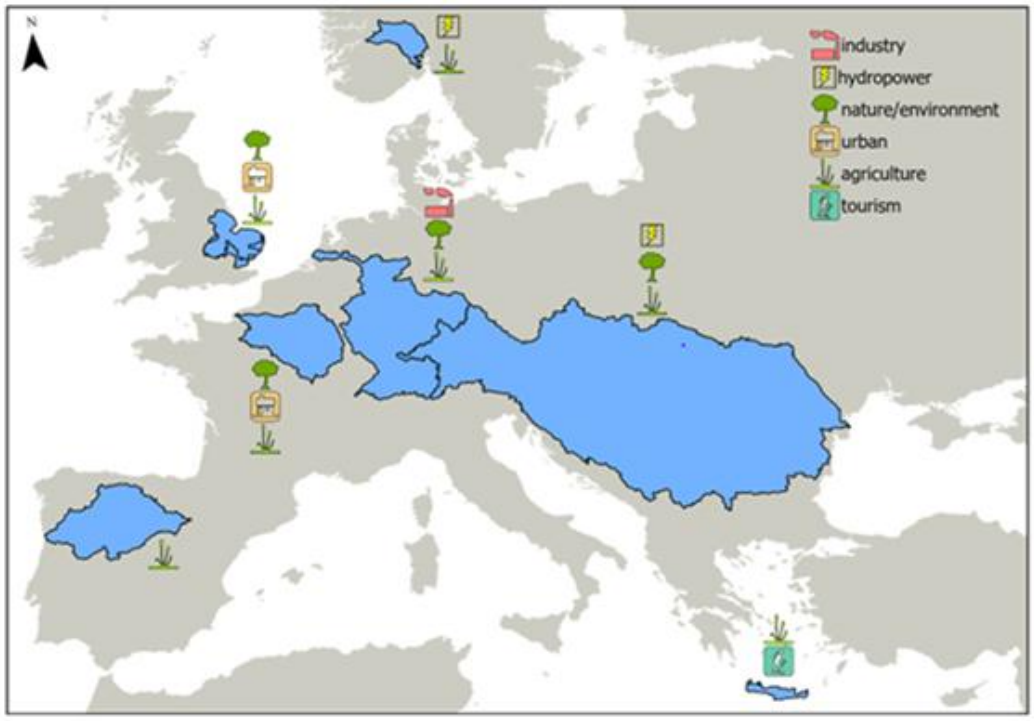


Figure 2. Overview of the seven river basin hubs of the STARS4Water project and the main sectors impacted by changes in water resources availability due to changing climate and socio-economic developments.

1.2 Work Package 3

STARS4Water Work Package 3 (WP3) is dedicated to the comparison, evaluation and further development of models for water resources management, which are key tools used to support decision making on water resources management all around the world. The objectives of this work package, entitled "Developing next generation models", are to:

- compare and evaluate existing models used operationally within the 7 RBHs and explore options for improving their accuracy and spatial resolution to meet stakeholder needs;
- test how innovative, multi-scale model integration and data-driven techniques could enhance water resources management.

The main activities in this work package are:

- benchmarking existing models against stakeholder requirements (Task 3.1);
- improving existing models based on new process understanding (Task 3.2);
- improving existing models based on multi-scale model integration (Task 3.3);
- developing data-driven water resources tools for selected river basins (Task 3.4).

The activities within this work package range across different themes related to water resources including groundwater, water quality, environmental flows and water demand. These themes were decided based on the stakeholder needs, which were gathered across the 7 RBHs via a series of local workshops and presented in report D1.2, "Assessment of the needs on data services and modelling tools of stakeholders in selected European river basins".

1.3 This report – D3.4

This report presents the data-driven modelling tools which have been developed in Task 3.4. This work has drawn on report D3.1: *Gap analysis of existing modelling tools in the STARS4Water river basins*, which details the models currently in use across the RBHs and, by comparing against stakeholder modelling requirements, identifies any modelling gaps or challenges which can be addressed in tasks 3.2, 3.3, or 3.4. In a series of workshops with the project team, a set of data-driven modelling tools were proposed based on the identified modelling gaps, as follows:

- Predicting reservoir storage and inflows: to address the need for better quantitative assessments of water resources, and to support decision making for reservoir operation and water resources management.
- Total water storage downscaling: downscaling satellite observations of total water storage to estimate monthly water table depth.
- Agricultural water use: to assess the impacts of climate change on agricultural water demand.
- Predictive mapping of groundwater quality: to address groundwater quality concerns in the RBHs.
- Quantitative groundwater resources estimation: to refine the estimation of groundwater resources, supporting water resources management and improving understanding of groundwater impacts on droughts and low flows.

In this report, each of these tools is described, and results of their application over selected RBHs are presented and discussed (full descriptions of each RBH can be found in report D1.1). A final section draws conclusions on these tools, reflecting on how they have met the stakeholder needs and the future potential for their wider application.

Data-driven Tools and Machine Learning

In recent decades, the unprecedented levels of accessible data and computational power has led to an explosion of applications of data-driven and Machine Learning (ML) models across almost all areas, including that of hydrology and water resources. There are many user-friendly generic and hydrology-specific ML packages available, reducing the barriers to application of ML for these purposes. Data-driven tools leverage the increasing volumes of data to: enhance existing models through hybrid approaches, data assimilation, and emulators; provide insights into system processes through feature importance and other explainable ML techniques; and add to our modelling capabilities through the development of ML models. Data-driven models have the benefit of assuming no prior knowledge as they learn patterns from the data provided, and they have the potential to capture non-linear relationships in the data which makes them well suited to modelling complex systems. Expert knowledge is then applied to refine and validate the model since, much like statistical methods, it is easy to produce misleading results if models are applied without care. Analogously to statistical methods, one key limitation to data-driven models is that they are generally not robust when applied outside the domain that they were trained on.

Each of the ML models described in this report are examples of supervised learning models, that is, they each see a set of explanatory variables with accompanying target variable(s) that they use to "learn" the relationship between the explanatory and target variable(s). This learning is then tested on a section of the data that has previously been withheld, to assess whether the model can accurately predict unseen target variables. A range of different ML models are explored in this work, including

tree-based models such as Random Forest and related algorithms, and neural-network type models such as Long Short-Term Memory (LSTM) and related algorithms. A brief summary of each of the modelling tools is provided in Table 1.

Table 1. An overview of the data-driven tools applied in this work alongside the river basin hub (RBH) they were developed/applied in.

Aim	Applied Models	Timestep & Prediction Horizon	Spatial Resolution & Extent	RBH	Input Data Type	Validation Data Type
Reservoir inflow and storage prediction	LSTM	Daily; 1 day ahead	point; multi-reservoir	Duero; Seine	in-situ; global[1]	in-situ
Reservoir storage forecasting	Ensemble-tree	Monthly; 1-3 months ahead	point; multi-reservoir	Duero; East Anglia	in-situ; global[1]	in-situ
Downscaling water table depth anomalies	RF; LSTM	Monthly; n/a	9 km; basin	Seine	satellite; reanalysis; simulated	simulated; in-situ
Irrigated area prediction	RF	Annual; n/a	1 km; basin	Rhine	satellite	in-situ (survey statistics)
Groundwater spatial contamination prediction	Ensemble-tree	n/a; n/a	~km; basin/aquifer	Duero; East Anglia	spatial[2]	in-situ
Groundwater storage change prediction	STT; XGBoost	Monthly; 1 month	11 km; basin	Duero	spatial[2]	simulated; in-situ

1. global catchment-level data, e.g. meteorological data and static catchment characteristics
2. spatial datasets, such as lithology, land-use and rainfall, which have been produced using a range of methods

2 Predicting reservoir storage and inflows

2.1 Introduction

Surface reservoirs are a major source of water for human use, and it is important to have a reliable forecast for reservoir status to ensure efficient operation of individual reservoirs and the wider water resource system (Peñuela et al., 2020; Ahmad & Hossain, 2019). This is a current and vital issue for many regions worldwide and will become increasingly important under changing climate and growing demand for water and hydropower (Gleick et al., 2013; Gleick, 2003). Surface water reservoirs have been identified as a modelling priority in four of the seven STARS4Water RBHs: Drammen, Duero, Messara, and Seine.

Given the importance of reservoirs in water resources, there has been a lot of research on predicting reservoir status at different lead-times and with different methods at a range of scales. Our interest is in data-driven methods, which have become popular as an alternative to hydrological models for forecasting. Data-driven methods have been used to forecast reservoir levels (Ibañez et al., 2021; Sapitang et al., 2020) which can be used to estimate reservoir storage or as a proxy for storage, and to forecast reservoir inflow (Yang et al., 2017; Hong et al., 2020), reservoir outflow (Yang et al., 2016), and reservoir storage anomalies (Tiwari & Mishra, 2019).

Basin stakeholders have highlighted the need for better quantitative assessments of water resources, and of the water stored in reservoirs. A modelling tool for both the monitoring and forecast of reservoir storage/levels was therefore put forward as high priority for the Duero, Messara and Seine basins, with the aim that such an approach could provide improved representation of reservoirs and operations, and as such aid the future planning of water allocation for the domestic, agricultural and energy sectors.

Here two methods are explored for reservoir prediction at different scales:

1. A Long Short-Term Memory (LSTM) model for simulation of reservoir inflow and storage at a daily timestep.
2. Ensemble-tree models (i.e. Random Forest (RF)-type models) for reservoir storage at a monthly timestep, designed to produce seasonal to sub-seasonal forecasts.

These methods differ both in their approach and their intended usage, but both address the need for improved quantification of water resources in reservoirs. An LSTM model at a daily timestep has the advantage of producing simulations for reservoir storage and inflow at a high temporal resolution. Therefore, the outputs are well suited to short-term decision making, with the added advantage of simulating reservoir inflow, which is useful information for water managers. Consequently however, this model has high data requirements which can be difficult to satisfy (i.e. historical daily timeseries of reservoir inflow and storage). In comparison, ensemble-tree models have lower data requirements, particularly since this application only requires historical reservoir storage at a monthly timestep. Thus, it is much easier to source data and to apply this model to multiple reservoirs. The ensemble-tree method is designed to produce forecasts on a monthly timestep at 1-3 months ahead, so the outputs can support decision making at those timescales. The disadvantage of this method is that the forecasts are at a coarse timestep and do not include reservoir inflow.

2.2 Long Short-Term Memory (LSTM) models for reservoir inflow and storage

One of the main challenges for establishing an LSTM-based model for reservoir inflow and storage predictions is the relatively low amount of available data. Hence, several approaches are explored for artificially enhancing the data, in addition to establishing a multi-task learning (MTL) approach for reservoir predictions (Figure 3).

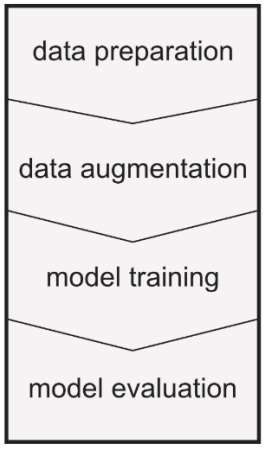


Figure 3. Workflow summary of the LSTM training.

Data augmentation: in machine learning, data augmentation is used as an umbrella term for techniques that increase the dataset size by creating modified versions of the original data to improve the learning process. The most common approach uses invariants, where the data augmentation procedure modifies the inputs, and the model must predict the unchanged labels. For instance, when we are interested in classifying digits from an image, we can implement a form of data augmentation by rotating the images. For a given sample, say a 7, we might rotate the input image by a small amount and let the model predict that the corresponding label is still a 7. Some data augmentation techniques require an intimate knowledge of the application domain, while others are largely domain independent (and can thus in principle be applied to all kinds of tasks). Within STARS4Water, we are mainly interested in the latter since individual reservoir storages can be managed substantially differently. Specifically, we explore four approaches: (1) noise addition observations, (2) scaling of the observation, (3) CMixup (Yao et al., 2022), and (4) Moving Block Bootstrap (MBB). We explain each of these approaches in more detail in the methods section.

Multi-task learning: in the machine learning domain, MTL refers to a set of approaches that involve training a model to perform multiple related tasks at the same time to improve generalization and efficiency for a task of interest. MTLs specifically utilise the LSTM architecture, which is well-suited for processing sequences of data with long-term dependencies as found in rainfall–runoff modelling. The resulting MTS-based LSTM is designed to learn from and predict multiple sequences of data simultaneously, each representing a different task (specifically: reservoir inflow and storage). This facilitates the exploration of regime change based on a 'transfer' of parameters between the different reservoirs. By doing this, there is potential to forecast reservoir storage for longer lead times, such as on seasonal scales. This approach might address some of the more fundamental drawbacks around ML approaches and their ability to forecast outside of the intrinsic boundaries of the input data.

Method

Multi-task learning: daily reservoir inflow, outflow and volume observation data from the river basin hubs, along with meteorological inputs and socio-economic statistics from HydroAtlas (Linke et al., 2019), are used to train an LSTM model to simulate naturalised streamflow and historic reservoir volumes. We use MTL for training an LSTM, i.e. the model is trained to simultaneously predict inflow and reservoir storage. However, the primary goal is reservoir storage, because it represents the component with limited data availability. The naturalised streamflow (which becomes the inflow to a given reservoir) serves as a secondary task. Henceforth, we refer to this as the MTL-based LSTM.

The resulting MTL-based LSTM model is specifically designed to learn from and make predictions across multiple sequences of data simultaneously, with each sequence representing a distinct task. This allows the model to share knowledge between related tasks. In our case, these tasks are the estimation of (naturalised) reservoir inflow and the respective storage. Our current model uses Caravan, a global community dataset for large sample hydrology (Kratzert et al., 2023), for training. The dataset includes meteorological forcings data, static catchment attributes (e.g. geophysical, sociological, climatological), and streamflow data for more than 6830 catchments across the globe. We split the dataset into training and validation sets for model selection. To assess the potential of MTL-based LSTM, we evaluate the model on the time series of five case study reservoirs in the Duero basin (Camporredondo) and the Seine River basin (Marne, Seine, Aube, and Pannecière). We also performed two different experiments: the first experiment without local inputs, which we named "simulation"; and the second experiment with local inputs (i.e., high quality input products from Great Britain), which we named "simulation (local products)".

Due to the high data demands of such an approach, fine-scale (<100 m) earth observation (EO) data from Planet (project partner) is also used to derive water surface area, and thereafter water level/volume, supplementing the data provided by the RBHs. Depending upon the results acquired from this initial model training, it is hoped to further explore the drivers of water use from different sectors, such as agriculture and energy. Initially, this would be done via proxies based on land use and energy prices, for example, and would again utilize EO data as provided by Planet and/or freely available EO data, where appropriate. More detail on this can be found in Appendix A: : Monitoring Embalsa Camporredondo using Planet Fusion.

Data augmentation: for the augmentation experiment we compare four techniques:

- *Additive Noise*: Gaussian noise is added to each timestep (with zero mean and two different standard deviations). The idea behind the input noise injection is that the modified inputs could generate new input patterns that mimic potential variations in the test data, therefore potentially making the model more robust to overfitting and able to handle real-world scenarios. In other words, the added noise introduces additional variability, which allows the model to learn from slightly altered versions of the same data points.
- *Scaling Noise*: a multiplicative noise factor is applied to the whole input. Here, the noise-scalar is drawn from a Gaussian distribution with mean equal to one.
- *Moving Block Bootstrapping (MBB)*: new samples are generated by resampling yearly blocks of sequential data (with replacement) (Künsch, 1989). It divides the time series into blocks or segments of contiguous data points. MBB generates new samples that maintain the statistical properties and temporal dependencies of the original data, helping to create a more diverse dataset for training machine learning models.

- *CMixup*: it samples close pairs of examples with higher probability and linearly interpolate based on label similarity. Details of the approach can be found in Yao et al. (2022).

We train and validate an LSTM for each augmentation technique and for each combination of parameters. All LSTMs have a hidden size of 256 (that is, the number of state that the model uses at each time step is 256) and a linear layer with a Relu (Rectified Linear Unit) activation function on top of it. We chose a dropout rate of 0.4 and Mean Squared Error (MSE) as a loss function. The learning rate for ADAM (an adaptive optimization algorithm (Kingma & Ba, 2017)) is 0.003 on the first 10 epochs (where an epoch is defined as one complete pass through the entire training dataset by the neural network during training), then 0.0005 from the 11th–15th epoch, and 0.0001 for the remaining 5 epochs. Our baseline is the LSTM trained on the original dataset, then compared against the models where we incorporate augmentation techniques.

We also experiment with the addition of local weather products in Great Britain. Specifically, the addition of high-quality input products from the CAMELS GB dataset. This addition is of scientific interest. The input has no direct influence of the predictability on the hydrological predictions in France and Spain. However, the conjecture is that high quality inputs for some parts of the data could have the potential to nudge the model to learn representations that generalise better.

Results and Discussion

Multi-task learning: the MTL-based LSTM is evaluated on the time series of the five case study reservoirs. We use the following evaluation metrics (table 2):

i) the Nash-Sutcliffe Efficiency (NSE), which measures the magnitude of the residual variance with respect to the variance of the observed values,

ii) the Kling-Gupta Efficiency (KGE), which computes the Euclidean distance between the relative variances, the bias and Pearson's correlation coefficient, and

iii) the Pearson's correlation coefficient (r), which measures the linear relationship between the observed and the predicted values.

Table 2. Evaluation for a Multitask Learning based Long Short-Term Memory model (MTL-LSTM) simulating daily reservoir inflow and volume, using Nash-Sutcliffe efficiency (NSE), Pearson's correlation coefficient (r) and Kling-Gupta efficiency (KGE). Comparison between the model trained with global input products and the model trained with global and local meteorological forcings from Great Britain (GB) in conjunction. The values in bold are the best performance.

Reservoir	MTL-LSTM (global)						MTL-LSTM (global and GB)					
	NSE		r		KGE		NSE		r		KGE	
	Inflow	volume	Inflow	volume	Inflow	volume	Inflow	volume	Inflow	volume	Inflow	volume
Pannecière	0.577	**0.832**	**0.854**	**0.98**	0.704	**0.835**	0.569	0.579	0.834	0.932	**0.738**	0.734
Marne	0.694	**0.942**	**0.852**	**0.974**	0.636	0.903	**0.721**	0.926	0.85	0.964	**0.764**	**0.913**
Seine	0.794	**0.952**	**0.902**	**0.981**	0.737	0.883	**0.81**	0.949	0.901	0.976	**0.834**	**0.916**
Aube	0.506	**0.97**	0.718	**0.99**	0.668	**0.946**	**0.606**	0.969	**0.781**	**0.990**	**0.73**	0.942
Camporredondo	**0.807**	-2.306	0.899	**0.397**	**0.842**	-0.114	0.790	-5.302	**0.918**	0.235	0.693	-0.196

The closer to 1 the values of NSE, KGE and r are the better the model. The results show that the LSTM can in general model both inflow and volume relatively well. The exception is the Camporredondo reservoir, where the modelling of the reservoir failed. Our hypothesis here is that the LSTM would require additional information to model the reservoir behaviour more accurately. The high metrics for the reservoir volume simulation are explained by the high seasonality of the volume. In practice, it remains difficult to capture details about systemic behaviour of reservoirs. The LSTM captures some large-scale processes, but remains unable to model long-term strategies (e.g., when reservoir operators strategically withhold water earlier in the year to release it at some later stage).

The addition of the "local products" (i.e., high quality input products from Great Britain) exhibits mixed results. Interestingly, the inflow simulations seem to generally improve under the local product addition. This can be seen as an indication that our conjecture that high quality inputs enabling the learning of better representations might hold. However, the improvement seems to be very specific and does not generalize to the volume estimations.

Figure 4 and Figure 5 show the hydrographs of daily inflow and storage observations compared with the simulations. The model performs well on each reservoir for the inflow and the storage ($0.56 < NSE < 0.81$, $0.83 < r < 0.91$), except the storage for the Camporredondo which provides the worst performance (NSE = -2.31, r = 0.40).

Data augmentation: for data augmentation, we find that when training the model on a low number of basins most augmentation approaches exhibit good performance across all evaluated metrics (

Table 3). The exception here is CMixup, which exhibits a lower prediction performance than the baseline. The best techniques we found are Scaling and MBB. However, the evaluation of the model trained on the 12187 basins (i.e., Caravan and GRDC-Caravan) shows marginal improvements over the baseline. This can also be seen from the associated distribution of the NSE values.

While all approaches, except for CMixup, lead to statistically significant performance increases, the increases are rather small from a hydrological standpoint. We assess the statistical significance (p) using the two-sided Wilcoxon (paired) signed-rank test, we also compute and report the corresponding effect sizes (using Cohen's d), and find the best performing approaches are Scaling Noise ($d=0.005$, $p=2e^{-6}$) and MBB ($d=0.051$, $p=2.6e^{-3}$). In summary, we found that none of the methods provide a strong basis for artificially augmenting the data. We repeated the experiment using CAMELS-US (Addor et al., 2017) (not shown here), and the results yielded a similar pattern

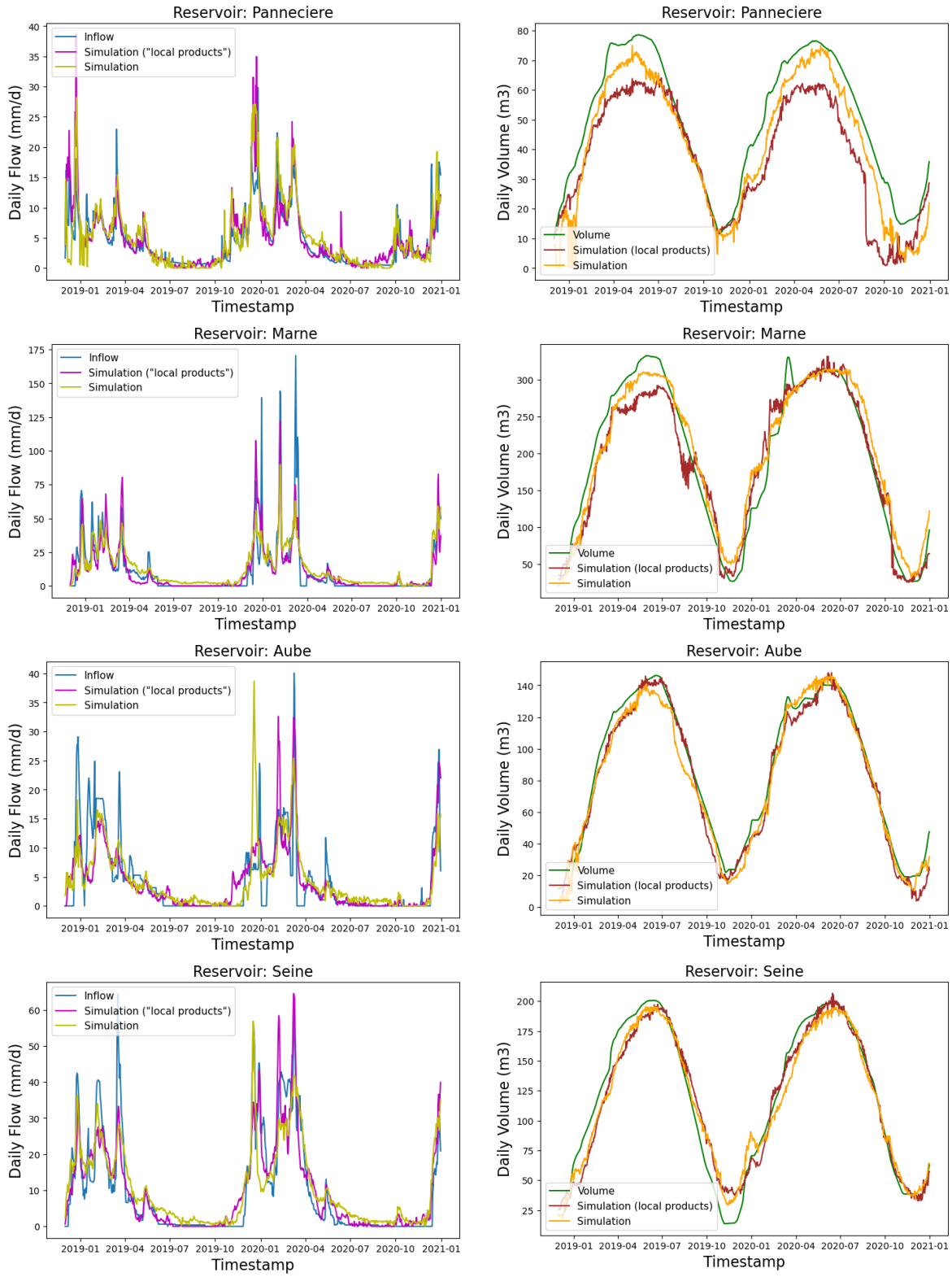


Figure 4. Hydrograph of daily inflow (left) and storage observations (right) compared with their simulations for the Pannecière, Marne, Aube, and Seine reservoirs. The simulations with the "local products" refer to the addition of high-quality data from Great Britain as inputs for model training.

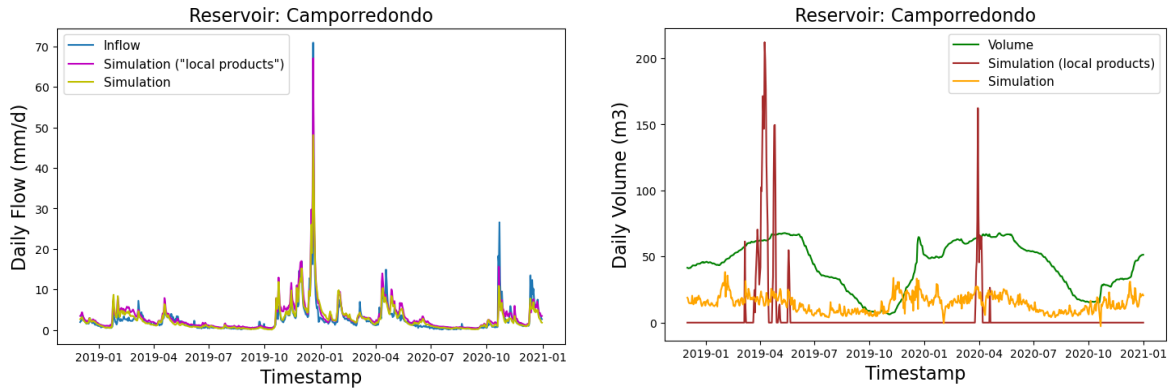


Figure 5. Hydrograph of daily inflow (left) and storage observations (right) compared with their simulations for the Camporredondo basin. The simulations with the "local products" refer to the addition of high-quality data from Great Britain as inputs for model training.

*Table 3. Comparisons of the evaluation metrics (Nash-Sutcliffe efficiency (NSE), Kling-Gupta Efficiency (KGE), Pearson's correlation coefficient (r)) between the baseline model and the data augmentation approaches using the inflow. Here, Noise means additive noise, scaling means multiplicative noise, CMixup is a mixing variant for continuous data and MBB stands for moving block bootstrapping (see methods). The model is trained and evaluated on 10, 50, and 100 basins. Statistics are averaged over different runs. Performances that are better than the Baseline are marked in **bold** and the best overall value in **violet**.*

Metrics	Aggregation	Baseline	Noise (std=0.03)	Noise (std=0.2)	Scaling	CMixup	MBB
10 basins							
NSE	Median	0.727±0.01	**0.749***±0.005*	**0.75**±0.003	**0.737**±0.013	0.694±0.009	**0.737±0.015**
	Mean	0.66±0.007	**0.667**±0.008	**0.67**±0.002	**0.662**±0.008	0.618±0.007	**0.662**±0.012
R	Median	0.868±0.005	**0.875***±0.003*	**0.876**±0.002	**0.871**±0.004	0.842±0.003	**0.874**±0.003
	Mean	0.839±0.001	***0.842****±0.001*	**0.839**±0.003	**0.839**±0.004	0.805±0.001	0.836±0.002
50 basins							
NSE	Median	0.77±0.006	**0.784**±0.003	**0.781**±0.007	**0.792**±0.007	0.763±0.011	**0.774**±0.006
	Mean	0.569±0.047	**0.621**±0.032	**0.587**±0.038	**0.593**±0.062	**0.619**±0.014	**0.549**±0.085
R	Median	0.895±0.004	**0.897**±0.003	0.893±0.002	**0.899**±0.004	0.87±0.022	**0.896**±0.002
	Mean	0.844±0.002	**0.847**±0.002	**0.844**±0.002	**0.847**±0.002	0.829±0.003	**0.849**±0.0
100 basins							
NSE	Median	0.763±0.007	**0.771**±0.002	**0.772**±0.009	**0.768**±0.012	0.733±0.019	**0.774**±0.006
	Mean	0.661±0.019	**0.64**±0.002	**0.772**±0.009	**0.768**±0.012	0.733±0.019	**0.774**±0.006
R	Median	0.887±0.004	**0.893**±0.003	**0.892**±0.003	**0.889**±0.004	0.877±0.1	**0.889**±0.001
	Mean	0.837±0.001	0.83±0.002	**0.837**±0.002	**0.837**±0.002	0.821±0.006	**0.839**±0.003

Conclusions and next steps

We examined how suitable an LSTM-based multi-task setting can be to model reservoir inflows and volumes on a daily timestep. While the inflow modelling shows quite robust and promising results, the performance of the volume predictions strongly depends on the reservoir operations. This should not be confused with the absolute value of the metrics (since the signals of the volume and the inflow are very different in nature). It is very likely that the reasons for the lack in generalizability are (a) missing inputs that describe the reservoir operations, and (b) the low amount of training data for the reservoir volume prediction. To address this scarcity, we explored different data augmentation approaches in a setting where we artificially increased the number of basins for training. Our data augmentation experiments indicate that predictions can be improved for low data regimes (<100 training basins) but not by much. For larger data regimes (>100 training basins) the augmentation approaches seem to get amortized: as more data becomes part of the training the improvements due to augmentation become marginal. This indicates that, for now, no task-agnostic data augmentation techniques exist that can expand large datasets in such a way that additional model capabilities are built. The results are in accordance with insights from the vision domain, where simple task-informed augmentation (say, mirroring an image from a bus from left to right, while maintaining the label “bus”) are heavily used (Shorten & Khoshgoftaar, 2019; Iwana & Uchida, 2021). Whether such augmentation can be found, constructed, and exploited efficiently for hydrological applications remains a question for future research.

2.3 Ensemble-tree models for reservoir storage

Method

Ensemble-tree models are explored as an option for predicting reservoir storage at a monthly timestep, in both simulation and forecasting mode. These models have the benefit of being less data-hungry and computationally expensive, more user-friendly, and with a higher degree of explainability compared to LSTM and other deep-learning models. These models are applied to reservoirs in the Duero basin and reservoirs across the UK (including the Anglian region).

This approach has been implemented in two ways: i) by building individual models for each reservoir using timeseries data specific to that reservoir, and ii) by building a multi-reservoir model which uses timeseries data for all the reservoirs included (Duero and UK) as well as reservoir and catchment characteristics sourced from the Caravan and CAMELS datasets (Kratzert et al., 2023; Delaigue et al., 2024).

Individual models maximise computational efficiency and flexibility by minimising the number of data types required but may be less accurate than a multi-reservoir model, particularly for extremes, since the training data will be more limited. Using a multi-reservoir model has shown promise for data-driven rainfall-runoff modelling (Kratzert et al., 2019), but comes at the cost of both increased data demand and training time. These two approaches will be compared here.

The data used for these models is presented in Table 4. The individual models only require time series data, whereas the multi-reservoir model requires all the data listed. It should be noted that some of the catchment characteristics have been calculated from timeseries over the catchment, and if those timeseries extend into our testing period (e.g. those calculated from the ERA5 dataset) we have recalculated them on a backwards-looking ten-year rolling window to ensure no data-leakage between the testing and training data.

The process of building an ensemble-tree model for reservoir storage prediction is summarised in Figure 6 and more details can be found in Appendix B. Four case study reservoirs in the Duero basin, described in Table 5 are used to explore the potential of ensemble-tree models for simulating reservoir storage at 1 and 3 month lead times (note that these case study reservoirs are not the same set used in the LSTM model). For the 3 month simulation, the model is run recursively, i.e. using the simulated storage for months 1 and 2 to predict month 3, but with observed precipitation and temperature variables. In forecast mode, these variables will be replaced with forecast precipitation and temperature.

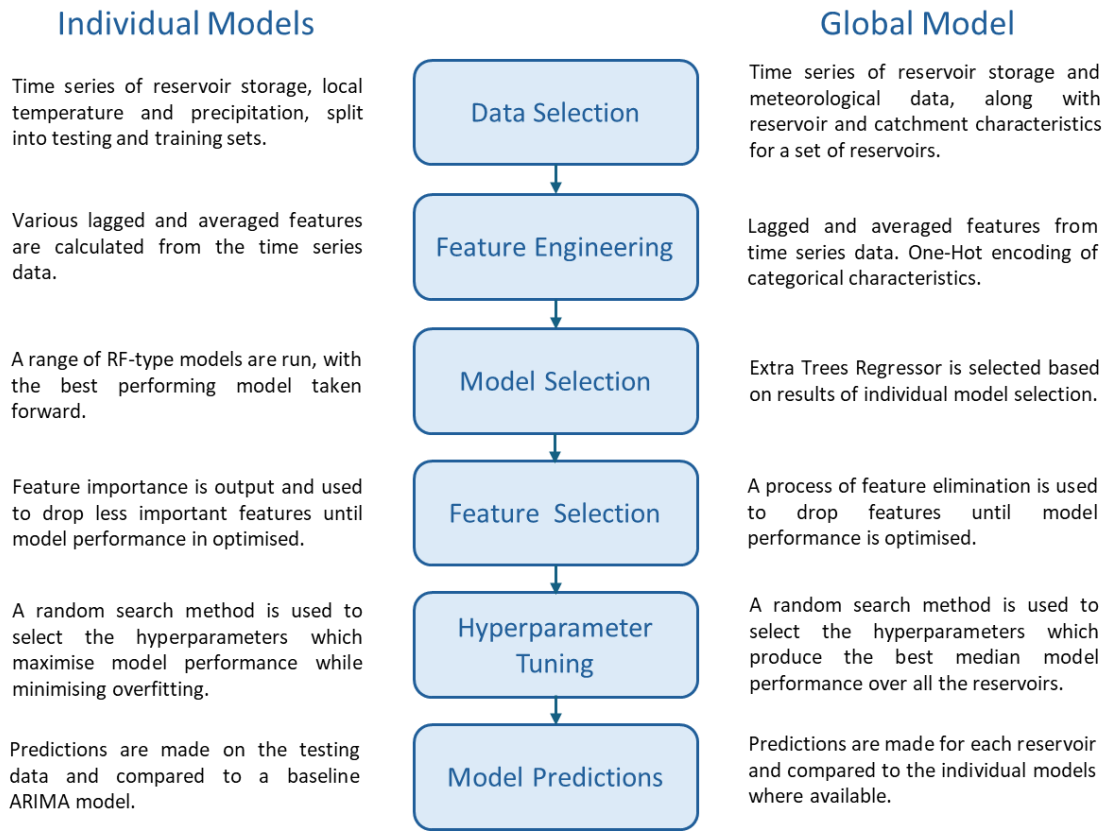


Figure 6. Summary of the steps taken to build individual Random Forest(RF)-type reservoir models, and the Extra Trees (ET) multi-reservoir model.

Table 4. Summary of the data used to build Extra Trees (ET) models for reservoir storage.

Time series Data	Data Source	Catchment Characteristics from HydroAtlas	Catchment Characteristics from ERA5 data	Data Source
Reservoir storage	(CEDEX, 2024; NRFA, 2022; Hollis et al., 2019)	Natural discharge[4]	Aridity index	(Kratzert et al., 2023; Delaigue et al., 2024)
Precipitation		Degree of regulation	Fraction of precipitation falling as snow	
Temperature		Elevation[5]	Average duration of high precipitation events[10]	
		Groundwater table depth	Frequency of high precipitation days[10]	
Reservoir Characteristics	**Data Source**	Inundation extent[6]	Average duration of low precipitation events[11]	
Total capacity	(Lehner et al., 2011; Durant & Counsell, 2018; Hughes M., 2004)	Limnicity - percent lake area	Frequency of low precipitation days[11]	
Catchment area		Lake Volume	Precipitation[8]	
Operator		Reservoir volume	Potential evapotranspiration[8]	
Impounding or non-impounding[1]		River area	Moisture index[9]	
Individual or grouped[2]		Land surface runoff	Seasonality	
Purpose[3]		Stream gradient	Air temperature[8]	
		Snow cover extent[7]		
		Land cover class		
		Human development index		
		Irrigated area extent (equipped)		
		Population count		
		Soil water content[8]		
		Urban extent		

1.Distinguishes between reservoirs with inflows primarily fed by streamflow and those by pumped water. **2**.Distinguishes between single water body reservoirs, and those made of linked water bodies (e.g. cascading reservoirs). **3**. e.g. water supply, irrigation, hydropower, flood control. Allowance is made for reservoirs with more than one purpose. **4**.Annual min/max/mean; **5**.min/max/mean; **6**.Annual mean/min, long-term max; **7**.Monthly mean, annual max/mean; **8**.Monthly mean, annual mean; **9**.Annual mean; **10**.High-precipitation defined as ≥5 times mean daily precipitation. **11**.Low-precipitation defined as <1 mm per day.

Table 5. Description of the case study reservoirs used to explore reservoir storage prediction with ML methods.

Reservoir Name	Latitude	Longitude	Capacity (hm^3)	Surface Area (km^2)	Catchment Area (km^2)	River
Camporredondo	41.89	-3.26	70	4.19	231	Carrión
Porma	42.93	-4.71	318	11.79	253	Porma
Santa Teresa	40.67	-4.40	496	27.19	1853	Tormes
Cuerda del Pozo	41.87	-1.30	249	22.10	550	Duero

Forecasting. To explore the potential for forecasting with these models, we test the multi-reservoir ET model in forecast mode over reservoirs in the UK, using monthly and seasonal ensemble meteorological forecasts derived from Historic Weather Analogues (HWA) and produced by the UK Met Office (Stringer et al., 2020) to drive the model. These meteorological forecasts are currently used to produce hydrological forecasts over the UK (UKCEH, 2025). The seasonal meteorological forecasts have 510 ensemble members, and here we use the periods: March, April, May (MAM); June, July, August (JJA); September, October, November (SON); and December, January, February (DJF). The monthly meteorological forecasts have 140 members.

The monthly forecast skill has been evaluated for UK reservoirs using the fair Continuous Ranked Probability Score (CRPS) (Wilks, 2011). The CRPS compares the cumulative distribution functions of the forecast to the observation for each month and is analogous to the squared error, with a perfect forecast giving a CRPS of 0. The 'fair' CRPS accounts for the finite size of the forecast ensemble by correcting the score towards that which would be obtained from an infinite ensemble (Ferro, 2014). This metric is then used to calculate the Continuous Ranked Probability Skill Score (CRPSS) (Wilks, 2011), defined as:

$$CRPSS = 1 - \frac{\langle CRPS_{forecast} \rangle}{\langle CRPS_{reference} \rangle}$$

where angled brackets denote an average of the variable within it. $CRPS_{forecast}$ are the CRPS scores of the monthly ensemble forecast, and $CRPS_{reference}$ are values for a reference forecast: in this case the distribution of reservoir storage for a given month is used. Therefore, the CRPSS score describes the *value added* by using the model over assuming reservoir storage follows historical patterns, with positive scores indicating added value and 1 an optimal forecast. Since the reference CRPS is calculated from a historical distribution of observed reservoir storage, only reservoirs with a sufficiently long record are used in this analysis.

Results

Model evaluation. Model performance on the test data is evaluated using NSE which has a score of 1 for perfect predictions. For each of the individual models, the Extra Trees Regressor (ET) was selected as the best performing model. Comparison between individual and multi-reservoir ET models are made for four case study reservoirs in the Duero: Camporredondo, Porma, Santa Teresa, and Cuerda del Pozo (described in Table 5). Model performance for storage simulations for the multi-reservoir ET model is presented for all the reservoirs included in the model (from the Duero and the UK), and forecast performance is evaluated for UK reservoirs only.

Comparison between simulation skill for the multi-reservoir and individual ET models and our baseline ARIMA models (see Appendix B) at 1 and 3 months, for each case study reservoir, is presented in Table 6. Time series of observed and predicted reservoir storage are shown in

Figure 7 (note that only the individual ET models have been included to prevent the plots becoming cluttered). Both the ET and ARIMA models perform better for Porma and Cuerda del Pozo reservoirs, which have smoother storage patterns, compared to Camporredondo and Santa Teresa. Skill decreases for each model for 3 month ahead simulation compared to 1 month ahead, with the ET models showing a smaller decrease in skill compared to the ARIMA model. The ET models show a small improvement on the ARIMA models for 1 month ahead forecasts and perform better than the ARIMA model at a 3 month lead time. The multi-reservoir ET model performs better than the individual ET models at 1 month lead time for all the case study reservoirs, and at 3 months it performs better for the Porma and Santa Teresa reservoirs, with similar performance for Camporredondo and Cuerda del Pozo.

Table 6. Model skill (Nash-Sutcliffe Efficiency) in predicting reservoir storage at 1 and 3 months for individual Extra Tree (ET) models (indv), the multi-reservoir ET model (multi), and the baseline ARIMA model.

Reservoir	1 month			3 months		
	Extra Trees		ARIMA	Extra Trees		ARIMA
	indv	multi		indv	multi	
Camporredondo	0.84	0.86	0.83	0.69	0.68	0.64
Porma	0.95	0.97	0.97	0.81	0.89	0.80
Santa Teresa	0.90	0.91	0.89	0.71	0.74	0.62
Cuerda del Pozo	0.95	0.97	0.95	0.85	0.86	0.70

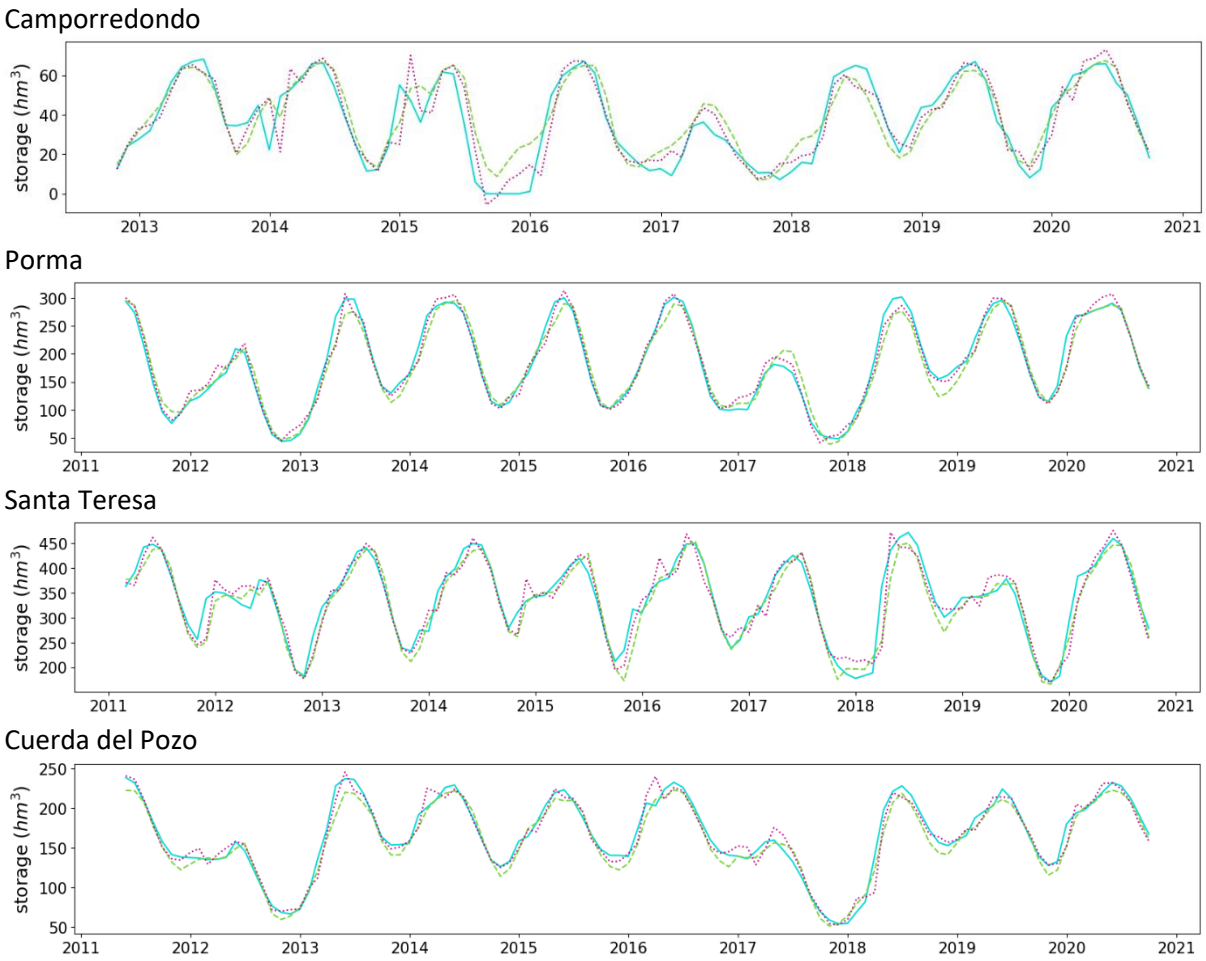


Figure 7. Reservoir storage: observed (solid blue); individual Extra Tree (ET) model (dashed green); ARIMA model (dotted pink), for 1 month ahead predictions.

The results for the multi-reservoir ET model are presented across all the reservoirs included in the model in Table 7. Here the model is evaluated for the whole test period (compared to the results presented in Table 6 where the evaluation period was cropped to match that of the individual models), so the results for the case study reservoirs are slightly different to those in Table 6. Predicted and observed storage for selected reservoirs from the multi-reservoir ET model are presented in Figure 8.

Table 7. Reservoir characteristics and model performance for the multi-reservoir Extra Tree model (Nash-Sutcliffe efficiency, NSE) at 1 and 3 months lead time.

Reservoir Name	Basin*	Country	Capacity (hm^3)	Catchment Area (km^2)	Purpose†	1 month NSE	3 months NSE
Barrios de Luna	Duero	Spain	308	499	Irr; WR; P	0.97	0.92
Porma	Duero	Spain	318	253	Irr; P	0.97	0.89
Riaño	Duero	Spain	651	593	Irr; P	0.97	0.89
Cuerda del Pozo	Duero	Spain	249	550	Irr	0.97	0.86
Villameca	Duero	Spain	20	56		0.97	0.91
Aguilar de Campoo	Duero	Spain	247	546	Irr; P	0.96	0.82

Linares del Arroyo	Duero	Spain	58	760	Irr; P	0.95	0.82
Roadford	South West	UK	35	31	WR	0.95	0.82
Bewl	Thames	UK	28	21	WR	0.94	0.80
Wimbleball	South West	UK	21	29	WR	0.94	0.88
Colliford	South West	UK	29	12	WR	0.93	0.80
Castro de las Cogotas	Duero	Spain	58	852	WR	0.93	0.72
Santa Teresa	Duero	Spain	496	1853	Irr; P	0.92	0.74
Elan Valley	Severn	UK	99	184	WR	0.92	0.86
Requejada la	Duero	Spain	65	221	Irr; P	0.91	0.77
Derwent Valley	Humber	UK	40	126	WR	0.91	0.83
Compuerto	Duero	Spain	95	308	Irr; WR; P	0.91	0.78
Uzquiza	Duero	Spain	75	150	WR	0.90	0.68
Ardingly	South East	UK	5	23	WR	0.90	0.74
Rutland	Anglian	UK	117	73	WR	0.88	0.63
Clatworthy	South West	UK	54	18	WR	0.87	0.81
Llyn Celyn‡	Dee	UK	70	60	WR	0.87	0.68
Vyrnwy	Severn	UK	55	74	WR	0.87	0.81
Camporredondo	Duero	Spain	70	231	Irr; P	0.86	0.65
Cervera de Ruesga	Duero	Spain	10	53	Irr; P	0.84	0.23
Arlanzon	Duero	Spain	22	104	Irr; WR	0.83	0.61
Loch Bradan	Scotland	UK	23	15	WR	0.83	0.65
Loch Katrine	Scotland	UK	111	194	WR	0.82	0.74
Daer	Scotland	UK	22	47	WR	0.82	0.71
Brianne	Western Wales	UK	62	88	WR	0.80	0.58
Llyn Brenig‡	Dee	UK	61	22	WR	0.48	-1.61

*In the UK, these are basin districts

†Irr: irrigation; WR: Water Resources; P: hydropower

‡Note that these reservoirs have shorter datasets so are used solely for model testing (i.e. not model training)

The multi-reservoir ET model shows good performance at all the reservoirs at 1 month ahead (NSE>0.8) with the exception of Llyn Brenig (NSE=0.48), although these high scores are due in part to the seasonality of the reservoir storage timeseries. Model skill decreases at 3 months, as anticipated, but remains high for most of the reservoirs (median NSE=0.78). In the multi-reservoir model, different reservoirs having different numbers of data points in the test and training set due to the train/test split (see Appendix B for details). Perhaps unsurprisingly, model performance across the reservoirs is correlated to the ratio of data points in the training set compared to the total number of data points for each reservoir. This also allows us to test the model on reservoirs that were not seen during the training phase, namely Llyn Celyn and Llyn Brenig.

Four reservoirs from the multi-reservoir ET model were selected for further investigation:

- Barrios de Luna, where the model performs well at 1 and 3 months lead time.
- Cervera de Ruesga, where the model performance significantly degrades between 1 and 3 months.
- Llyn Celyn, an 'unseen' reservoir where the model performs well.
- Llyn Brenig, an 'unseen' reservoir where the model performs poorly.

Time series of these reservoirs (observed and simulated storage) are shown in Figure 8. Visual inspection of this figure, highlights that the model performs very well at Barrios de Luna at both 1 and 3 month lead time, with some underestimation of the storage at high levels but good representation of the peaks in the drier years 2012 and 2017. The model performs reasonably well at Cervera de Ruesga for 1 month ahead predictions. However, it does overestimate storage during the drier winter months, but performance drops significantly at a 3 month lead time. Early data from Cervera de Ruesga (not shown) shows that the storage is very erratic during the initial ten years, though it does settle into a more regular pattern after 1990, which may contribute to the reduction in model skill. The model performs well at Llyn Celyn, despite the non-typical storage pattern. The model captures the historic low in the summer of 2022 with only a slight overestimation, although the 3 months ahead prediction has a tendency to overestimate storage during previous dry summer/autumn periods such as those in 2014 and 2018. At Llyn Brenig the model shows only moderate performance at 1 month, and poor performance at 3. The storage pattern at Llyn Brenig is very a-typical, which likely results in the poor model performance.

Forecasting. The CRPSS for UK reservoirs forecasting at a monthly timestep are presented in Table 8. Results show added value over the historical distribution for all reservoirs (i.e. CRPSS>0) for monthly forecasts and are roughly correlated with model performance in simulation mode.

Figure 9 and Figure 10 show an example of the reservoir storage forecasts at a seasonal and monthly timestep respectively, along with the meteorological forecasts used to drive the model, at the Rutland reservoir for selected years. The historic range of storage values for a given month or season is plotted, using the same percentile ranges as the UKCEH hydrological outlook (UKCEH, 2025): the 13th, 28th, 72nd, and 87th percentiles, which have been calculated on an expanding window. These examples allow us to explore whether the reservoir model is skilful in forecasting reservoir storage when the Rutland reservoir storage is unusually low (2011) and unusually high (mid to late 2012).

Seasonal forecasts at Rutland (Figure 9) show that the model is capable of forecasting reservoir storage at seasonal timesteps with reasonable precision when the storage levels are in a normal to notably low range: the observed values often fall within the ensemble of forecasts during these periods, and the median of the forecast is generally in the same category as the observed value. However, the model fails to capture the exceptionally high levels of storage in the JJA and SON seasons in 2012, with unexpectedly high rainfall and low temperatures likely contributing to the model performance.

Monthly forecasts (Figure 10) show a similar pattern, although the observed value falls within the ensemble of predictions less often, likely due to the smaller ensemble size (140 compared to 510) and the higher levels of variability at a monthly timestep. The model shows stronger performance during normal to low storage periods (2011 to March 2012) compared to high storage periods (April to December 2012). While the underestimation of storage in April 2012 can be attributed to a remarkably wet April that year, the model continues to underpredict for a number of months after that despite being updated with the previous months observed storage value.

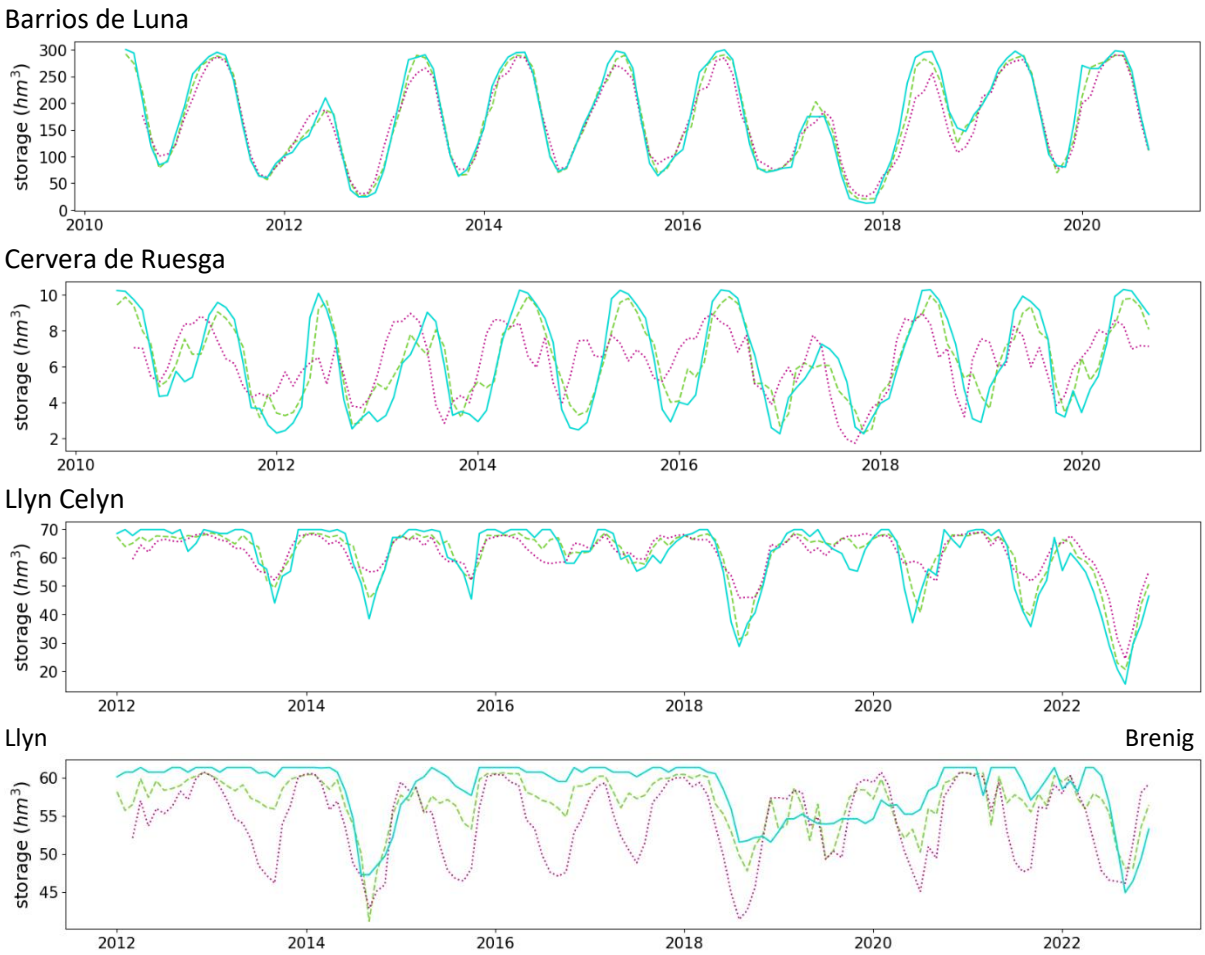
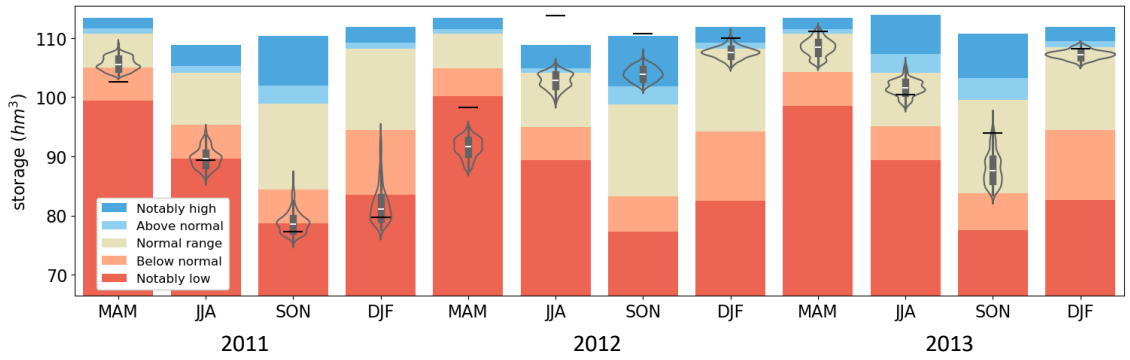


Figure 8. Reservoir storage for selected reservoirs in the multi-reservoir Extra Tree model: observed (solid blue); 1 month ahead (dashed green); 3 months ahead (dotted pink).

Table 8. Continuous Ranked Probability Skill Score (CRPSS) metric for UK reservoirs run in forecast mode at a monthly timestep. Only reservoirs with a sufficient historical record have been included.

Reservoir Name	Basin District	Capacity (hm^3)	Catchment Area (km^2)	Purpose†	CRPSS (monthly)
Roadford	South West	35	31	WR	0.76
Bewl	Thames	28	21	WR	0.67
Wimbleball	South West	21	29	WR	0.66
Colliford	South West	29	12	WR	0.82
Elan Valley	Severn	99	184	WR	0.37
Derwent Valley	Humber	40	126	WR	0.51
Ardingly	South East	5	23	WR	0.51
Rutland	Anglian	117	73	WR	0.56
Clatworthy	South West	54	18	WR	0.40
Vyrnwy	Severn	55	74	WR	0.37
Loch Bradan	Scotland	23	15	WR	0.49
Loch Katrine	Scotland	111	194	WR	0.28
Daer	Scotland	22	47	WR	0.30
Brianne	Western Wales	62	88	WR	0.23

a) Seasonal storage forecast, Rutland



b) Seasonal meteorological forecast, Rutland

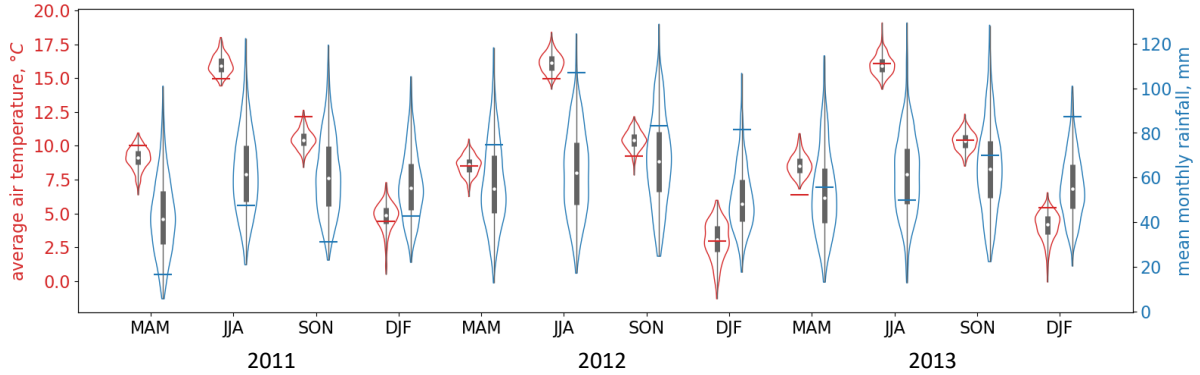
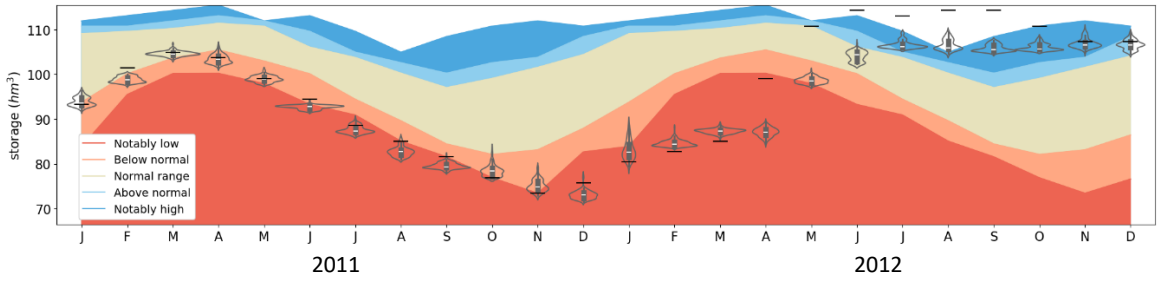


Figure 9. Seasonal forecasts at the Rutland reservoir: a) reservoir storage, b) rainfall (blue) and temperature (red) used to drive the model, for the years 2011-2013. Violin plots show the distribution of the ensemble members, a solid dash shows the observed values for that season, and the stacked bar lines indicate percentiles for the historic reservoir storage data for that season (calculated on an expanding window).

a) Monthly storage forecast, Rutland



b) Monthly meteorological forecast, Rutland

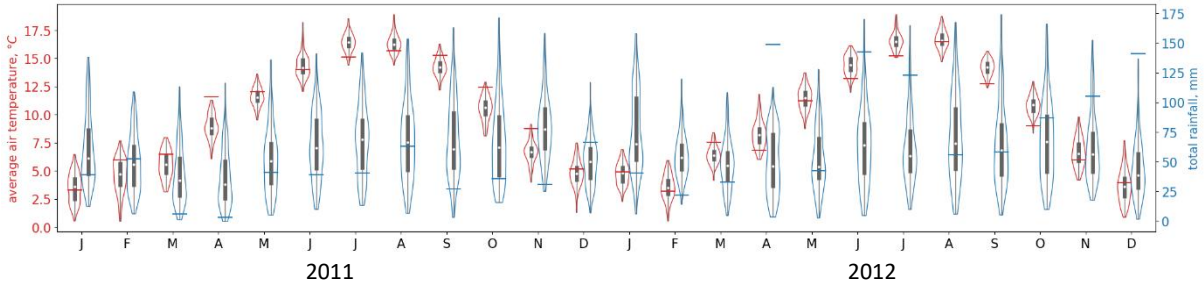


Figure 10. Monthly forecasts at the Rutland reservoir: a) reservoir storage, b) rainfall (blue) and temperature (red) used to drive the model, for the years 2011-2012. Violin plots show the distribution of the ensemble members, a solid dash shows the observed values for that month, and the stacked line plot indicates percentiles for the historic reservoir storage data for that month (calculated on an expanding window).

Discussion

Individual ET models show reasonable skill at predicting monthly storage in the selected case study reservoirs using historic storage and rainfall. Furthermore, they are quick to train and run, so have potential to be used for short-term forecasting driven by rainfall forecasts. They show similar performance to the baseline ARIMA models at 1 month lead time and outperform the baseline at a 3 month lead time.

The multi-reservoir ET model generally performs better than the individual models at both 1 and 3 month lead time for the case study reservoirs, so should be considered as an alternative to the individual models where possible. However, the multi-reservoir and individual models were only compared at four case study reservoirs, and a more thorough comparison would be required to confidently state that multi-reservoir models are preferable for this problem.

The data demand for the multi-reservoir model is significantly higher than the individual models, which can be limiting. As a counterpoint to this, the multi-reservoir model can be applied to reservoirs with shorter historic datasets, since the remaining reservoirs provide a substantial set of training data. This is demonstrated in the application of the multi-reservoir model to reservoirs that were not seen in the training data at all, although the performance of the model was mixed for these reservoirs, so further investigation would be useful.

Forecasting. The multi-reservoir ET model shows promise for use in forecast mode, with all reservoirs considered showing added value for the monthly forecast compared to using historical storage as a forecast method. Several years of monthly and seasonal forecasts were examined at the Rutland reservoir, and this analysis demonstrated that the model performed better in normal to low storage periods compared to high storage periods for the Rutland over the years 2011 to 2013. This could, in part, be due to the unexpectedly high levels of rainfall and low temperatures in the late spring and early summer of 2012 but may also be a result of the unprecedentedly high reservoir stores that were observed for those months. While high reservoir storages are present in the training data (many reservoirs in the training data regularly reach full capacity), the normal range for Rutland reservoir in the summer months is much lower than the 2012 values, so the model (which is trained on historical patterns) may have struggled to anticipate such values.

This snapshot analysis, while interesting, is far from comprehensive and much more analysis is required for this model in forecast mode.

Conclusion and next steps

Both the individual and multi-reservoir ET models show potential for predicting reservoir storage at a monthly timestep. The multi-reservoir model generally simulates reservoir storage well across a range of reservoirs at 1 and 3 months, and it would be useful to add more reservoirs into the model to explore the model potential in other regimes (e.g. the reservoirs in the Seine which are operated for flood control). It would also be good to compare more individual models to the multi-reservoir model to be able to give a more informed comparison.

Initial analysis of the multi-reservoir ET model in forecast mode has been undertaken, with promising results shown for reservoirs in the UK driven by the Met Office's HWA forecast. Future steps include a more thorough analysis of the forecast skill in the UK and extending the forecasting to reservoirs in other basins.

2.4 Overview of reservoir models

The challenge of simulating reservoir storage has been explored using two methods with distinct ambitions: an LSTM model to simulate daily reservoir inflow and storage, and ensemble tree methods to forecast monthly reservoir storage at 1 to 3 months ahead. The LSTM model demonstrated promising results for inflow simulations, and for storage simulations in four of the five reservoirs considered, but failed to accurately simulate storage for the Camporredondo reservoir (Duero). Scarce data on reservoir operations and low volumes of training data overall may have contributed to this result. Various data augmentation methods were trialled to address this, but only small improvements in model performance were seen. Therefore, this model can potentially be used to simulate reservoir inflow and storage on a daily timestep but may not be easily generalizable to new reservoirs, due to the data requirements and the possible performance issues.

The ensemble-tree models have shown good performance for simulating reservoir storage for most of the reservoirs modelled, with the multi-reservoir model generally outperforming individual reservoir models due to the increased amount of training data available to the multi-reservoir model. Initial results show good performance of the multi-reservoir model in forecast mode over the UK reservoirs. Although more analysis is required, the model demonstrates potential to be used as part of operational hydrological forecasts.

3 Estimation of monthly water table depth anomalies based on GRACE, ERA-5 and TSMP simulations

3.1 Introduction

Groundwater stores represent almost a third of global total freshwater resources and are an essential resource for sustaining many agricultural, industrial and domestic activities. In recent years, stresses on groundwater resources have been increasing due to the population growth. In addition, climate change might affect the natural recharge cycle of groundwater reservoirs by altering the precipitation and evapotranspiration patterns (Agarwal et al., 2023). Therefore, a suitable monitoring of groundwater levels (GWLs) and storages is essential to assess its potential and long-term sustainability. Improved groundwater models have been identified as research priorities for the Duero, East Anglia, and Seine river basins.

Traditional approaches to GWL assessment rely on the use of monitoring wells, providing observations for specific locations. However, measurements of the GWL taken at a single location might not be representative of the situation in the entire area of interest. Since in-situ measurements are not sufficient to assess continuous groundwater conditions, the use of numerical models has become an important tool for understanding groundwater dynamics over large areas (Lall et al., 2020).

Significant progress has been made in groundwater modelling, which plays an important role in the development and management of groundwater resources (Gaur et al., 2011; Akter & Ahmed, 2011). As the physical properties and processes governing groundwater flow are highly heterogeneous, many groundwater management problems require complex and fully distributed models that can well represent hydraulic properties with multiple boundary conditions. In addition, Kollet and Maxwell (2006) and Rahman et al. (2019) argue that surface-groundwater interactions should be incorporated into groundwater modelling to provide more reliable predictions. Thus, there has been a common interest to include more physically based models with higher spatial resolution up to continental domains (Condon et al., 2021).

Satellite remote sensing can also complement existing monitoring networks and modelling studies, filling gaps in spatial and temporal coverage. In particular, the NASA's Gravity Recovery and Climate Experiment (GRACE) can reliably measure monthly groundwater storage (GWS) change over large scale-areas (~200,000 km^2). This is useful to evaluate global total water storage (TWS) changes and their impacts associated with extreme event conditions and climate change. GRACE satellites have also been integrated with hydrological and land surface models to analyse TWS and GWS changes across large areas, such as the Amazon basin (Chen et al., 2009), the Yangtze River basin (Zhou et al., 2017), northwestern India (Rodell et al., 2009), and California's Central Valley (Famiglietti et al., 2011; Liu et al., 2019). However, due to its coarse spatial resolution, GRACE data cannot be directly used to investigate TWS changes in small catchments or local areas.

Given that GRACE has limitations at local and regional scales due to its coarse resolution, downscaling techniques are required (Wilby & Wigley, 1997; Atkinson, 2013). While computational processes of dynamical downscaling are more complex and require extensive computational time and resources, statistical downscaling methods have been receiving more attention to obtain high-resolution hydrological and climate data. Such examples include soil moisture (Wen et al., 2019; Peng et al., 2017), precipitation, temperature (Fasbender & Ouarda, 2010), and evapotranspiration (Tan et al., 2019). The statistical downscaling methods are easy to implement, and the downscaled results are

deemed of sufficient accuracy. They mainly establish a relationship between the low-resolution target data and variable data and then input high-resolution variable data into the regression model to output high-resolution target data (Chen et al., 2021).

This study aims to bridge the scale gap between coarse-resolution observations and high-resolution simulations for monitoring GWLs over large areas. It integrates GRACE satellite data, ERA5-Land reanalysis datasets, and simulations from the Terrestrial Systems Modelling Platform (TSMP) to estimate monthly water table depth anomalies (WTDA) in the Seine River basin. Two data-driven models, Random Forest (RF) and Long Short-Term Memory (LSTM) networks, were compared to emulate TSMP WTDA outputs and downscale GRACE data for local and regional groundwater management. The methodology involves constructing RF and LSTM models at pixel scale to simulate monthly water table depth time series at a spatial resolution of 0.11 degrees (TSMP-G2A). The results of this study are published in Avila et at (2025), this publication provides further details about the methodology.

3.2 Materials and Methods

Study area

The Seine River Basin (SRB), located in northern France, covers approximately 76,000 km² and lies within the sedimentary Paris Basin, a major European groundwater reservoir (Figure 11). It has a pluvial oceanic climate with an average annual precipitation of 666 mm and a mean annual discharge of about 600 m³/s at its outlet (F). The basin features interconnected aquifers in various geological formations, with groundwater significantly contributing to river flow (Rousset et al., 2004)). Land use is primarily agricultural (51%), followed by woodland (25%) and grassland (18%) (Mignolet et al., 2007). Annual water demand in the basin is about 1.8 billion m³ (23 mm/year per unit area), with 55% sourced from groundwater for drinking, industry, and irrigation (Tavakoly et al., 2019). While France has not experienced extreme aquifer depletion, long-term sustainability faces challenges from reduced recharge due to climate change, sea level rise, and future changes in groundwater usage (Maréchal & Rouillard, 2020).

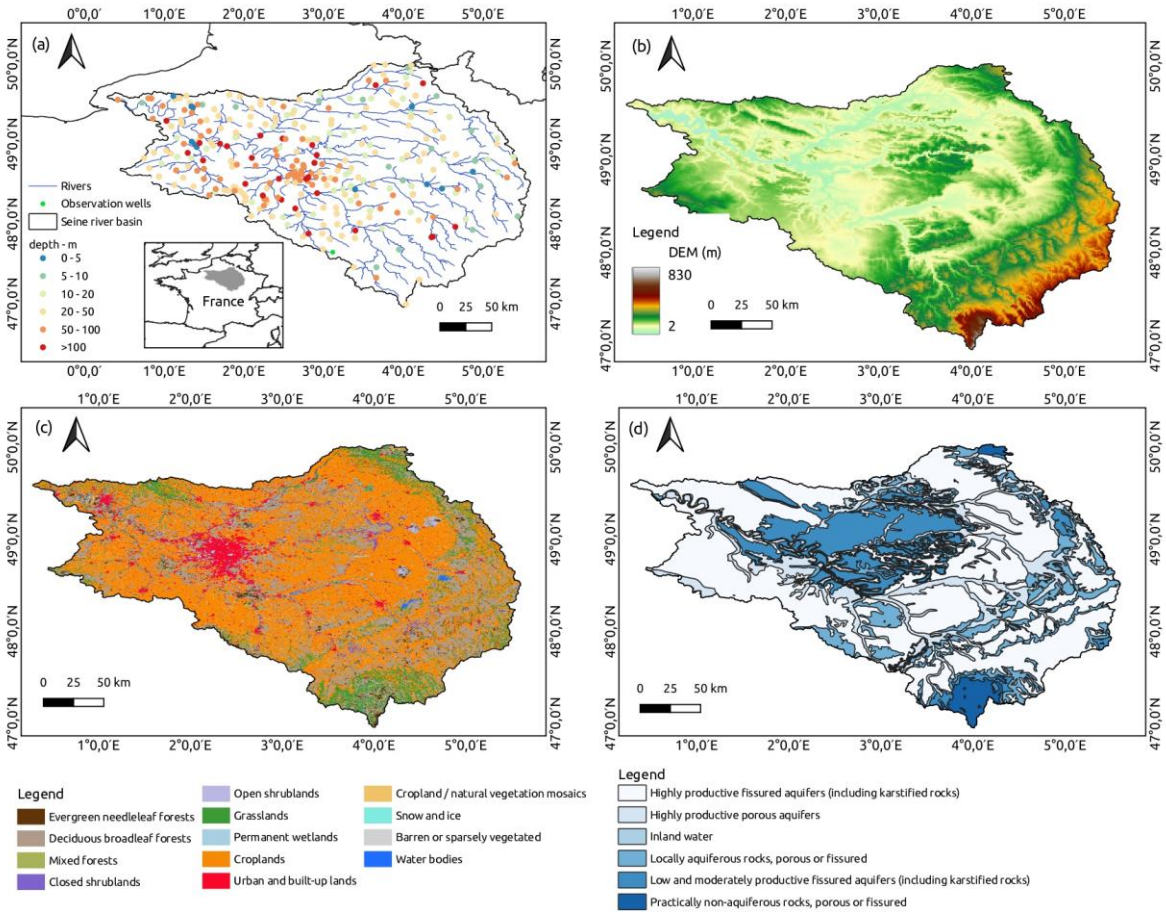


Figure 11. The Seine River Basin (a) Location and distribution of groundwater observation wells; (b) Digital Elevation Model (DEM); (c) Land cover; (d) Aquifer types.

General Approach

The adopted methodology to estimate monthly WTDA is based on the integration of large-scale hydrological simulations with remote sensing and reanalysis datasets. The models are trained at the pixel scale over the Seine River Basin, incorporating hydrological and climatological variables from ERA5-Land and TSMP-G2A datasets (Figure 12). The approach aims to emulate TSMP-simulated WTDA and refine GRACE data for local groundwater estimation. Model performance is evaluated by comparing results with TSMP simulations and in-situ groundwater observations, and the performance is assessed using Pearson correlation (r), Kling-Gupta Efficiency (KGE), and Root Mean Square Error (RMSE).

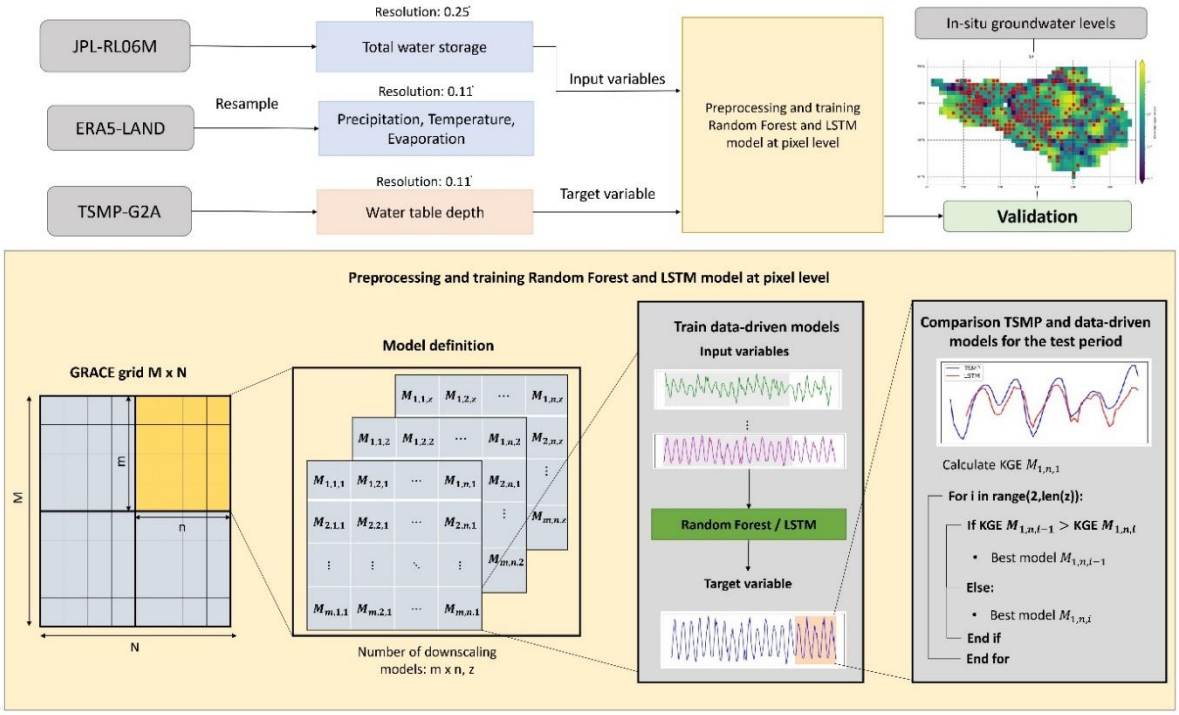


Figure 12. Framework of GRACE TWSA downscaling based on Terrestrial Systems Modelling Platform (TSMP) and Long Short-Term Memory (LSTM) network. The model considers Jet Propulsion Laboratory Release 06 (JPL-RL06M) of the GRACE data processing and ERA5-Land data to obtain the main meteorological variables.

GRACE Total Water Storage (TWS)

The GRACE (Gravity Recovery and Climate Experiment) satellites, jointly developed by the National Aeronautics and Space Administration (NASA) and Deutsches Zentrum für Luft- und Raumfahrt (DLR), operated from March 2002 to October 2017. Currently, GRACE solutions are mainly divided into two categories: Spherical Harmonic Coefficient (SHC) and Mass Concentration solutions (Mascon). For this study we used the GRACE TWS product (RL06 V1.0) from the Jet Propulsion Laboratory (JPL) mascon solution with a 0.5° × 0.5° grid. The JPL mascon approach models Earth's gravity field as discrete mass concentrations on a grid rather than as spherical harmonics. The JPL mascons are constrained by a priori information from geophysical models, such as global hydrological models, ocean circulation models, and satellite altimetry data (Watkins et al., 2015), and a coastline resolution improvement (CRI) filter is applied to reduce spatial leakage from land to oceans (Wiese et al., 2016). Temporal data gaps (~20% of total values) during the study period were interpolated using the spline method.

Reanalysis Data – ERA5-Land

The ERA5-Land reanalysis dataset provides high-resolution (9 km) global land variable data spanning from 1950 to near-present. It is generated through global high-resolution numerical integrations of the European Centre for Medium-Range Weather Forecasts (ECMWF) land surface model, utilizing downscaled meteorological forcing from the ERA5 climate reanalysis and incorporating an elevation correction for near-surface thermodynamic conditions (Muñoz-Sabater et al., 2019). This finer resolution, compared to 31 km for ERA5, enhances its suitability for applications in water resource management, land use, and environmental monitoring. Thus, it provides researchers and practitioners more accurate and detailed data for analysis and decision-making (Xie et al., 2022). For this study, hourly data from 2002 to 2022 were used to analyse air temperature at 2 m above the surface, daily

precipitation, and evaporation. Maximum and minimum monthly temperatures, as well as total monthly precipitation and evaporation, were estimated. The dataset is accessible at https://cds.climate.copernicus.eu.

Terrestrial System Modelling Platform (TSMP)

TSMP is a fully coupled atmosphere-land-surface-subsurface modelling system that gives a physically consistent representation of the terrestrial water and energy cycle from the groundwater via the land surface to the top of the atmosphere. TSMP integrates the numerical weather prediction model COSMO (version 5.0.1), the land surface model (CLM3.5) and the surface-subsurface hydrologic model Parflow (version 3.2). TSMP has been applied in many studies to simulate the terrestrial hydrological processes (Shrestha et al., 2014; Kurtz et al., 2016; Sulis et al., 2018; Keune et al., 2019), including the PRUDENCE (Prediction of Regional scenarios and Uncertainties for Defining European Climate change risks and Effects) regions. Furusho-Percot et al. (2019) evaluated the performance of TSMP simulations over Europe (TSMP-G2A dataset), indicating a good agreement of the hydrometeorological variability with different observed datasets at the regional scale in the PRUDENCE regions. They mainly compared anomalies of temperature and total column water storage with commonly used reference observational datasets, resulting in r ranging from 0.73 to 0.94 for temperature anomalies and from 0.62 to 0.88 for precipitation anomalies.

3.3 Results and Discussion

An exploratory analysis reveals a strong spatial correlation between WTD and input variables, particularly highlighting the influence of meteorological and local characteristics (Figure 13). While TWSA, maximum monthly temperature (Tmax), and evaporation exhibit well-defined seasonal patterns, precipitation shows greater variability without a consistent trend. The WTD variations across regions indicate that factors such as topography, soil type, and water use play a significant role. In the Seine basin, TSMP effectively captures groundwater conditions, showing a strong negative correlation with GRACE-derived TWSA (mean r=-0.58, median r=-0.68), suggesting that WTD increases as water storage anomalies decreases. This correlation supports the potential for machine learning models to estimate WTD anomalies using GRACE and TSMP data. Areas with deeper WTD (>10m) show weaker correlations, indicating reduced coupling with surface processes. Among ERA5-land variables, temperature exhibits the strongest positive correlation with WTD, followed by evaporation, while precipitation shows a weaker, mostly negative correlation, with sign reversals in regions with higher WTD values.

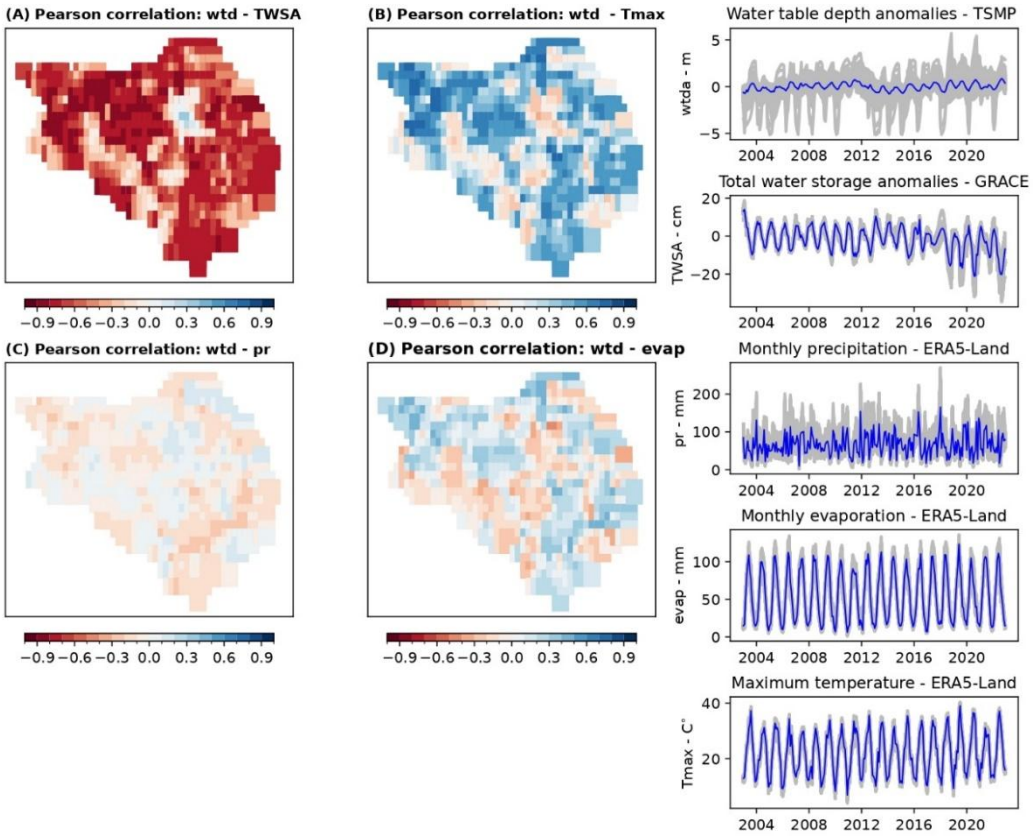


Figure 13. Spatial correlations and monthly time series for selected variables from GRACE and ERA5-Land with water table depth (WTD) anomalies obtained from TSMP. The blue time series represent the spatial mean, and the grey time series depict the variation across individual pixels.

Figure 14 presents the performance analysis of RF and LSTM models against TSMP simulations. Based on the KGE, r, and RMSE, the results indicate that LSTM slightly outperforms RF, with an average KGE of 0.54 compared to 0.52. The RF models achieve a lower average RMSE (0.1 m) than LSTM (0.23 m), while their r values are similar (0.61 for RF and 0.60 for LSTM). Lower KGE values are observed in regions where GRACE and TSMP show weak correlation, and RMSE analysis suggests lower errors in areas with shallower WTD. The time series comparison reveals that regions with shallow WTD (near the surface and main rivers) display a clear seasonal pattern, whereas deeper WTD regions exhibit more variability.

The independent validation of LSTM and RF models was conducted using in situ WTD observations, with TSMP-derived values as a baseline (Figure 15). Standardized values were calculated for 486 WTD wells, which were matched and/or averaged to the nearest pixel, resulting in 236 pixels with observed data. The r analysis, shown in Figure 15, indicates an average correlation of 0.3 across all wells, with a maximum of 0.89. Performance varied by location, with wells in the upper catchment (points d, e, and f in Figure 15) exhibiting stronger seasonal patterns and higher correlations (0.42–0.89) with lower RMSE values (0.99–0.28). These results suggest that groundwater dynamics in these regions are more closely linked to surface processes, leading to better model performance.

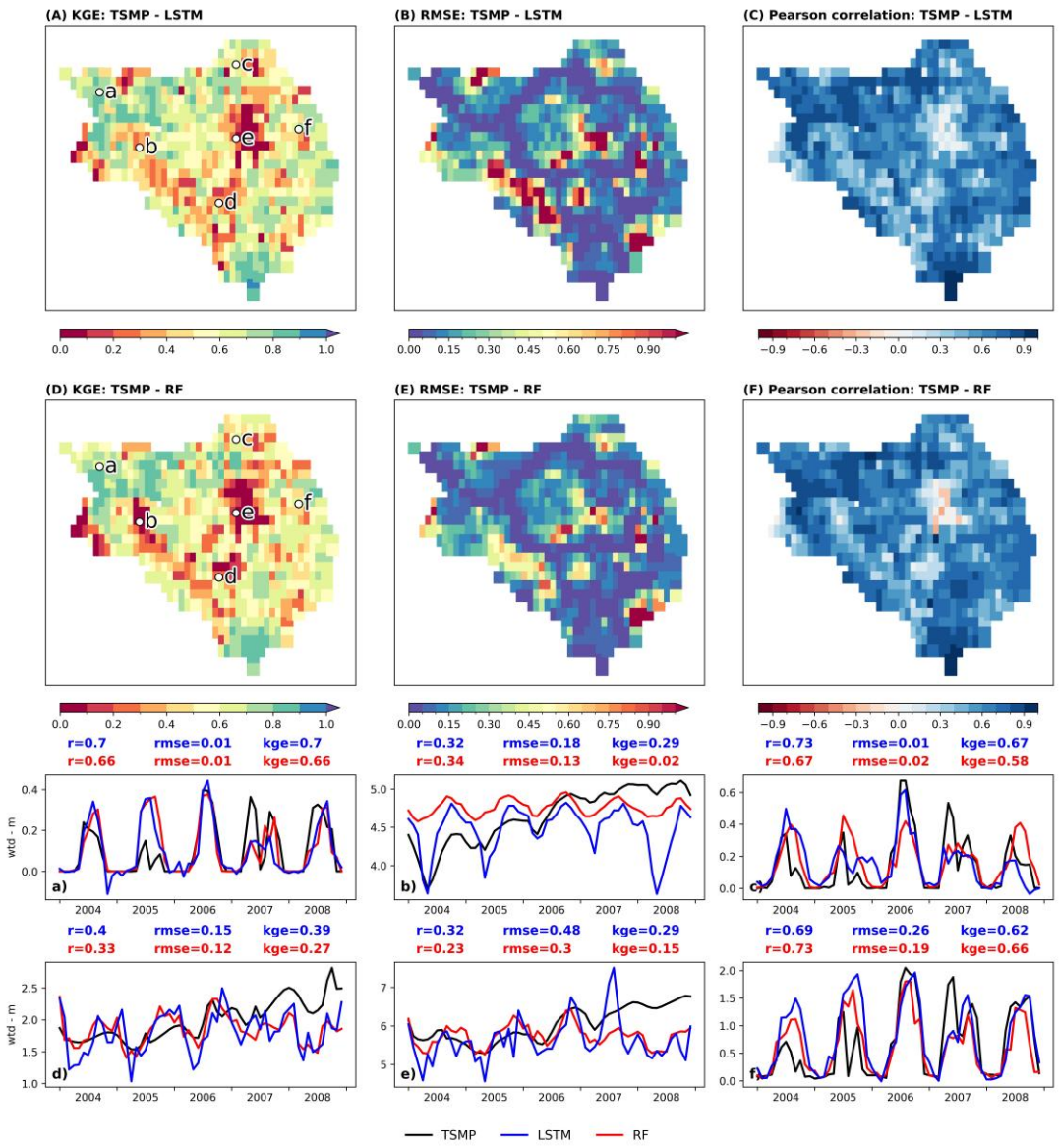


Figure 14. Comparative performance metrics of water table depth (WTD) obtained from Terrestrial Systems Modelling Platform (TSMP) and the downscaled WTD obtained from Long Short-Term Memory (LSTM) and random forest (RF) models. The metrics include: Kling-Gupta efficiency (KGE), root mean square error (RMSE) and Pearson's correlation coefficient (r).

Levene's test was conducted to assess the statistical differences between simulated and observed WTD time series at each well (Figure 16). On average, in 38% of the wells the null hypothesis was rejected at a significant level of α=0.01, while 55% did not reject at α=0.05. Here, α refers to the significance level which is the threshold probability for deciding whether to reject the null hypothesis. Additionally, 8% of the wells had p-values between 0.01 and 0.05. Like prior results, the downstream part of the basin exhibited more discrepancies between simulated and observed data, with orange dots mostly clustered in the centre.

Interpolated maps were also generated employing r and RMSE to create confidence maps for monitoring groundwater anomalies (Figure 17). Both the LSTM and RF models displayed similar spatial patterns, with lower correlations in the southern part of the basin and values as high as 0.7 in the northern part. While both models showed higher correlations in some areas, they did not significantly outperform the original TSMP simulations, as expected. The LSTM model, however, was less

consistent, with several regions having correlations below 0.2, unlike the RF model. Although the LSTM performance could be improved with further tuning (such as adjusting epochs and batch size), the RF model provided a better fit for simulating WTD anomalies at the pixel level and was more efficient overall. Given the multi-year memory between river discharge and precipitation, the lookback period should be extended, especially for LSTM models.

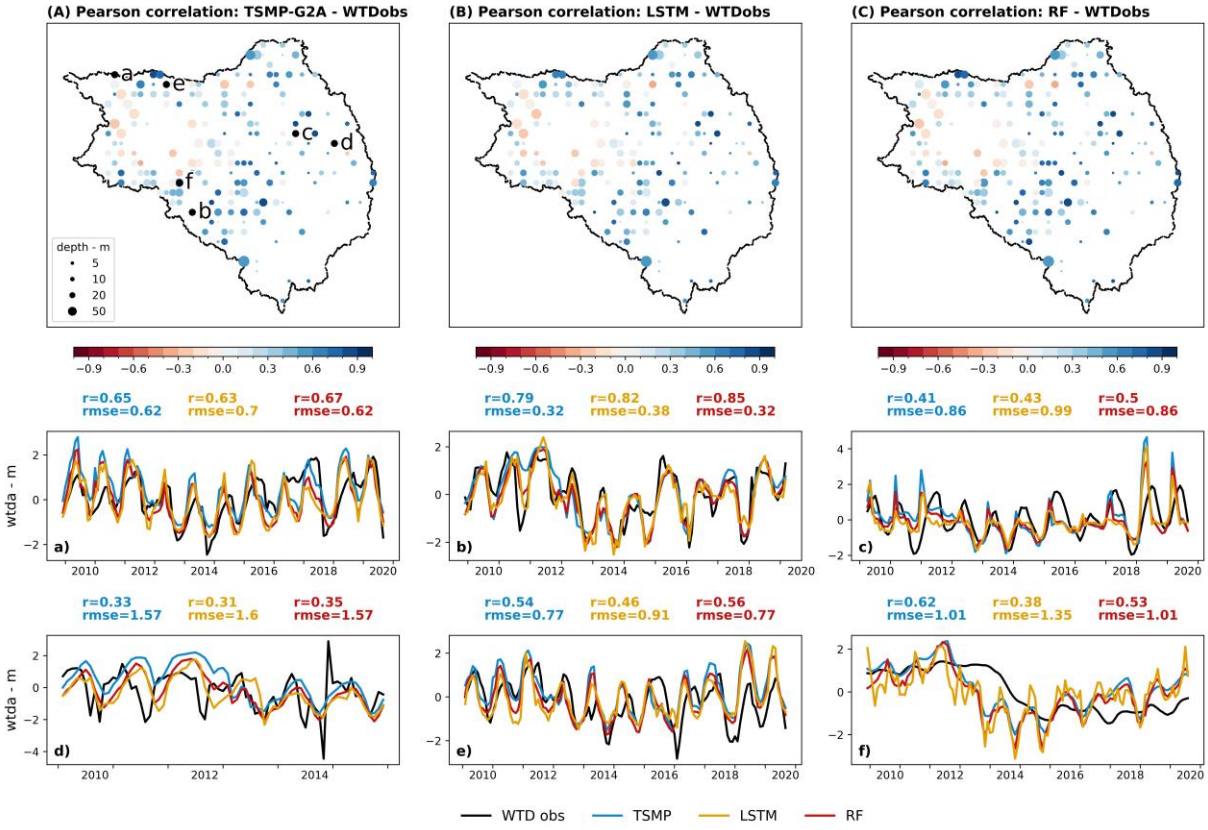


Figure 15. Pearson correlation coefficient between observed monthly water table depth anomaly and (A) that simulated by Terrestrial Systems Modelling Platform (TSPM-G2A) and the downscaled results obtained with (B) Long Short-Term Memory (LSTM)networks and (C) random forest (RF) for the period 2004–2018. (a)-(f) Time series for six arbitrary selected grid cells marked on the map in (A). Performance metrics including Pearson's correlation coefficient (r) and root mean square error (RMSE) are indicated above the time series plots.

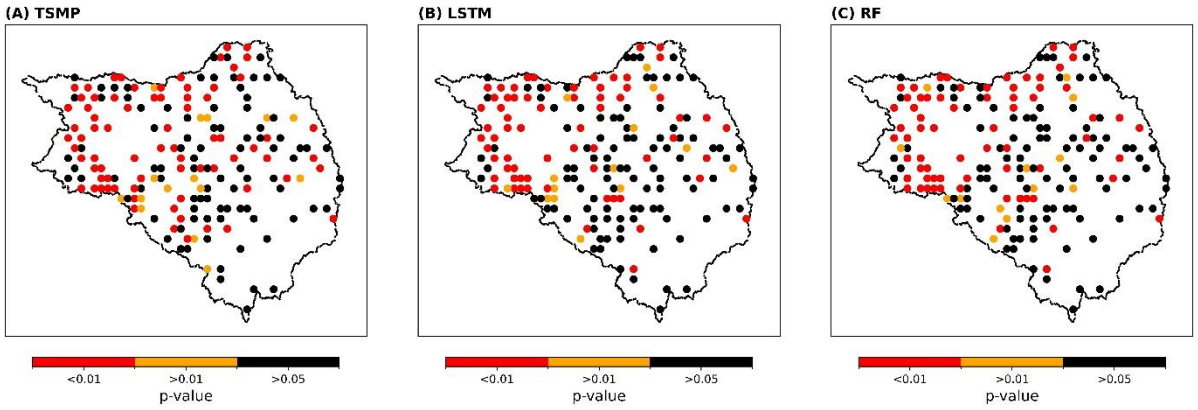


Figure 16. P-values obtained from the Levene test for the baseline Terrestrial Systems Modelling Platform (TSMP-G2A) dataset (A) and each data-driven model: Long Short-Term Memory (LSTM, B) and random forest (RF, C) at each water table depth observation well.

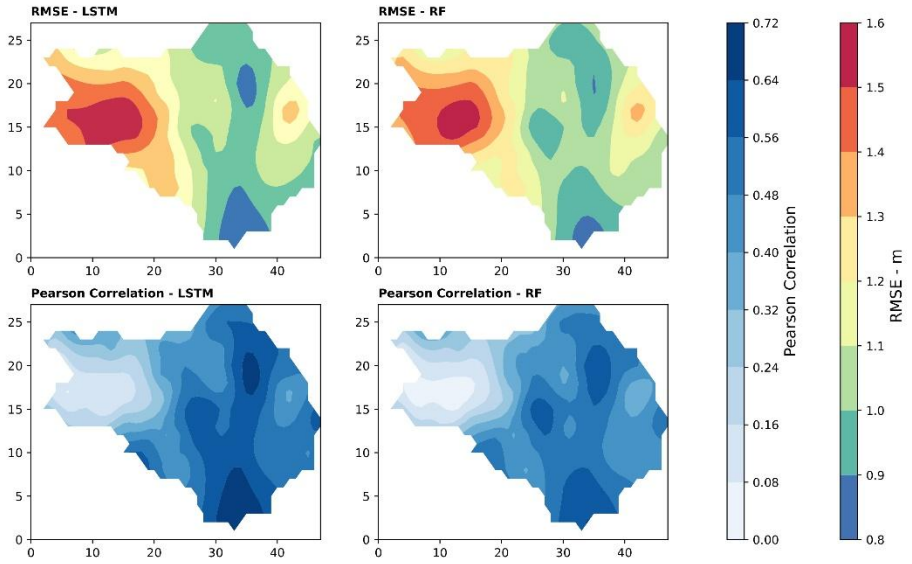


Figure 17. Interpolated root mean square error (RMSE) and Pearson correlation maps between observed and simulated groundwater depth anomalies over the Seine River basin. Simulations include results the downscaled results obtained with Long Short-Term Memory (LSTM) networks (left) and random forest (RF) (right).

3.4 Conclusions and next steps

This study demonstrates that RF and LSTM networks can successfully emulate complex hydrological simulations while enabling the downscaling of global satellite-derived water storage data. This offers a computationally efficient alternative to traditional physical modelling approaches. The pixel-scale modelling strategy proved particularly valuable by allowing region-specific optimization and flexible adaptation to local hydrological conditions. However, there were evident limitations in areas influenced by coastal processes and karst systems where global datasets lack sufficient resolution. Both RF and LSTM models performed comparably in capturing temporal water table dynamics. RF is more computationally efficient than LSTM and therefore may be more practical for operational applications. However, the potential for improved performance for the LSTM with extended temporal windows warrants further investigation. The discrepancies between model outputs and field observations in certain areas highlight the need for hybrid approaches that integrate local hydrogeological data with global datasets to enhance accuracy. These findings underscore machine learning's growing role as a complementary tool in hydrological sciences, capable of bridging the gap between large-scale climate products and local water management needs, while also pointing to important directions for future research including multi-objective learning frameworks, spatially explicit model architectures, and validation across diverse hydroclimatic regions. The study ultimately positions data-driven modelling as a valuable component of modern hydrologic assessment, particularly for rapid scenario testing and regional-scale analyses where traditional physical models may be prohibitively resource-intensive. This can address stakeholder requirements for high-resolution information on water table depth at low computational cost.

4 Agricultural water use

4.1 Introduction

As the climate becomes warmer and drier, understanding the historical and present inter-annual variability in agricultural water demand and water availability could provide insights into the potential pressures exerted by agricultural water demand on future water availability. Modelling of agricultural water demand has been identified as a research priority in the Rhine basin. Sustainable water resource planning and management can be improved by better understanding water consumption through the identification of irrigation areas and changes over time. There is an increasing need to expand such research efforts to support informed water management decisions. Changes in climate may benefit the agricultural industry in humid and temperate regions as more land becomes suitable for cultivating crops, leading to an increased need for irrigation water.

Most studies on impact modelling in agriculture have employed coarse spatial resolution when evaluating future global water availability and net irrigation requirements. Döll and Siebert (2002) applied the WaterGAP model at a spatial resolution of 0.5° to evaluate future global irrigation and water use under climate change. However, the spatial resolution used is considered insufficient to capture fragmented irrigated areas that significantly contribute to regional water use. A recent pan-European study utilized the FAO crop model for climate impacts modelling and the Inter-Sectoral Impact Model Intercomparison Project phase three (ISIMIP3) data have improved understanding of irrigation water requirements at coarse spatial resolution (Busschaert et al., 2022). The authors highlighted that the future increase in net irrigation water demand correlates with year-to-year variations in precipitation and atmospheric evaporative demand, however irrigated area itself was not explicitly included in their analysis.

As a first step towards estimating agricultural water demand, this study focuses on identifying the spatial extent of irrigated areas in the Rhine basin. The approach uses an ML model that integrates land surface temperature from MODIS observations and a hydrological model to improve irrigated area estimates. This work is based on our previously published study and aims to improve estimates of regional irrigation water demand at fine spatial resolution by first quantifying current irrigated areas (Purnamasari et al., 2025). Understanding the extent and distribution of irrigated cropland provides critical information for projecting future water needs under changing climate conditions. In this work, we apply a ML model to estimate irrigated area in the Rhine basin, integrating outputs from a hydrological model with land surface temperature observations.

4.2 Material and Methods

Study Area

The proposed method was tested for irrigated area in the Rhine basin which covers an area of approximately 160,000 km^2 (Figure 18). Agricultural land use occupies 46% of the total land use based on Copernicus Corine Land Cover (CLC 2018) (Figure 18b). An important characteristic of agriculture in the Rhine basin is the presence of irrigation systems in the Rhine Valley, located in the southern part of the basin along the French–German border. Agricultural activity is most prevalent from April to

September, when supplementary irrigation is sometimes required to prevent crop failure or improve yields (Purnamasari et al., 2025).

During spring and summer, the evapotranspiration rate exceeds the precipitation rate. These conditions translate to deficit precipitation to supply crop water requirements. The interannual variability of precipitation affects the total extent of irrigated area throughout the basin. Data on the total irrigated area in the Rhine basin at the sub-national level (NUTS 2 units) is available from Eurostat and presented in Figure 18c.

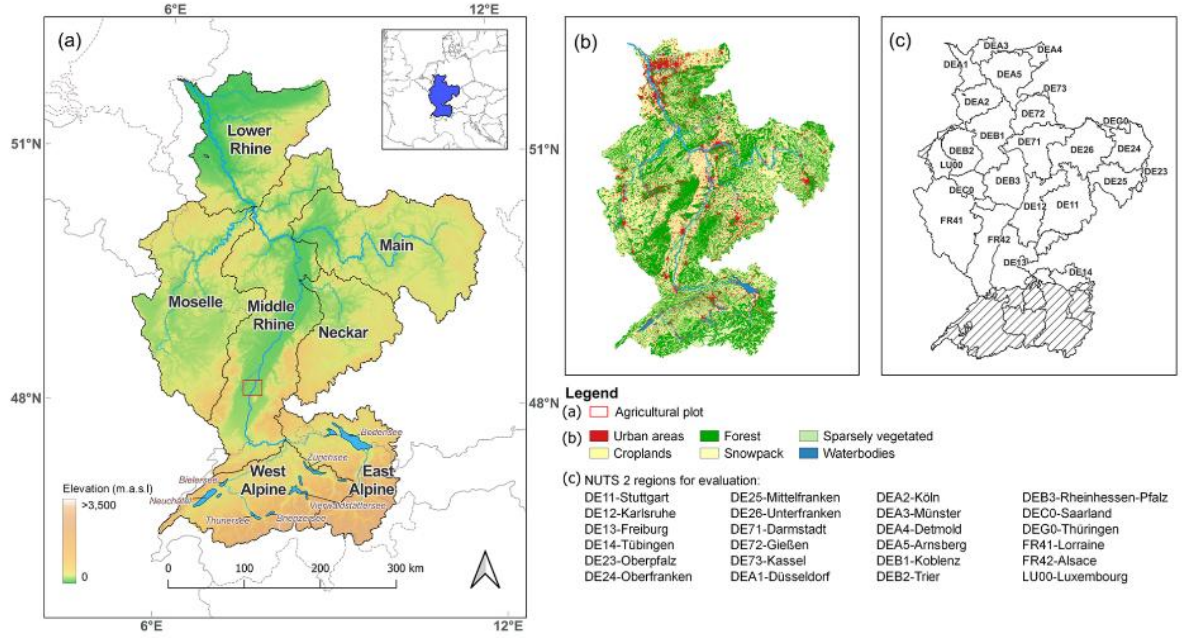


Figure 18. The overview of (a) subbasins, (b) aggregated land use and land cover according to CLC 2018, and (c) NUTS level 2 regions for which the reported irrigated area was used to evaluate the classification results are shown. The red box in panel (a) highlights the croplands used to collect training and test data for the supervised classification. Hatched regions in panel (c) indicate areas with no available reported data (Purnamasari et al., 2025).

Estimating Irrigated Area

The multiyear irrigated area at 1 km spatial resolution from 2010 to 2019 is defined using a supervised classification approach based on a random forest algorithm that distinguishes irrigated from non-irrigated land based on the daily spatiotemporal signature of land surface temperature (LST) differences (ΔT_s). Random forest algorithm is considered robust for spatial classification because its ensemble learning approach reduces overfitting and effectively handles complex spatial patterns by incorporating diverse spatial features. These ΔT_s are calculated by comparing observed temperatures from MODIS sensors onboard Terra and Aqua satellites ($T_{s,\,obs}$) (Wan et al., 2021) with simulated temperatures derived from actual evapotranspiration estimates produced by the spatially distributed hydrological model wflow_sbm ($T_{s,\,sim}$). The results are presented in Figure 19.

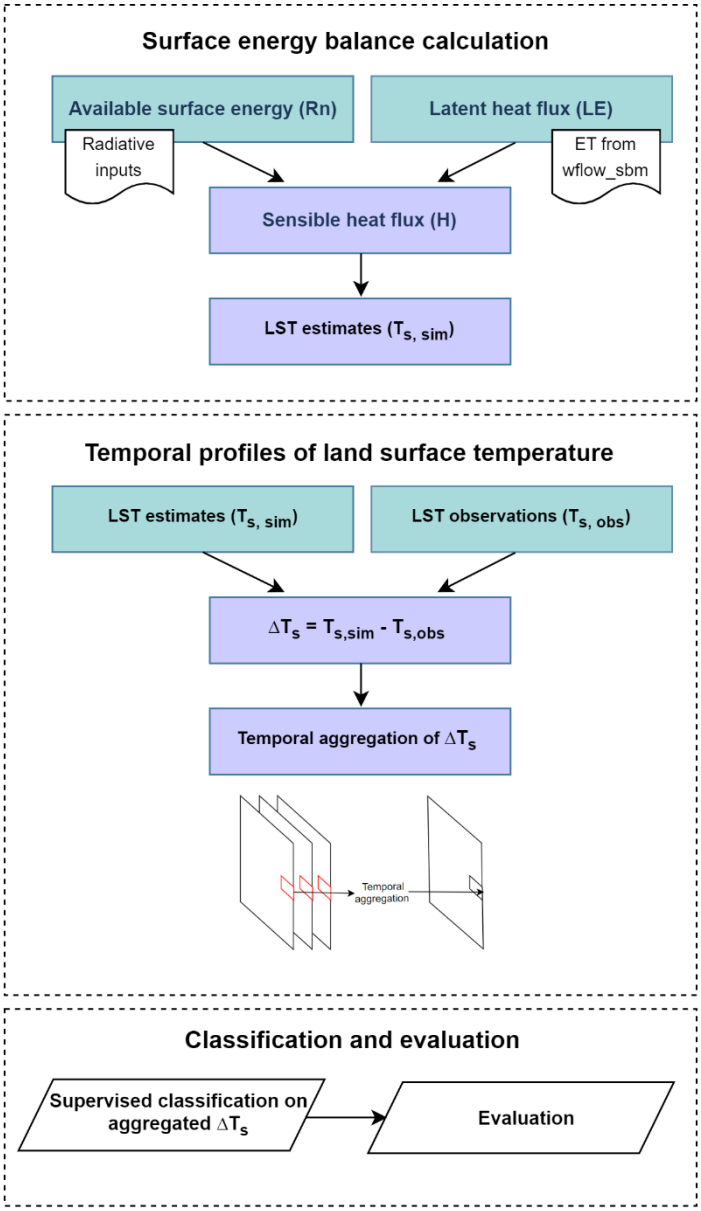


Figure 19. Overview of the methodology to identify irrigated area by using evapotranspiration estimates and land surface temperature observations (Purnamasari et al., 2025).

Training and testing data

After obtaining daily ΔT_s for the Rhine basin, we applied RF classification to produce consistent multiyear irrigated area. However, no pre-existing training and test datasets were available for the ΔT_s data. Therefore, high-resolution Landsat 7 and 8 imagery was used to generate point labels for the RF classification trained on ΔT_s data (Figure 20). True-colour images at 30 m resolution were used to visually identify irrigated agricultural plots with irrigation infrastructure, while 100 m thermal images helped reduce subjectivity of visual observations. To differentiate irrigated areas using thermal imagery, the standard deviation of land surface temperature over the growing season was calculated as irrigation tends to reduce temperature fluctuations. The dataset obtained from high-resolution Landsat 7 and 8 imagery was divided into two subsets: 80% as training data and 20% as test data. The training and test dataset for each year from 2010 to 2019 were collected to build an RF model for the corresponding year

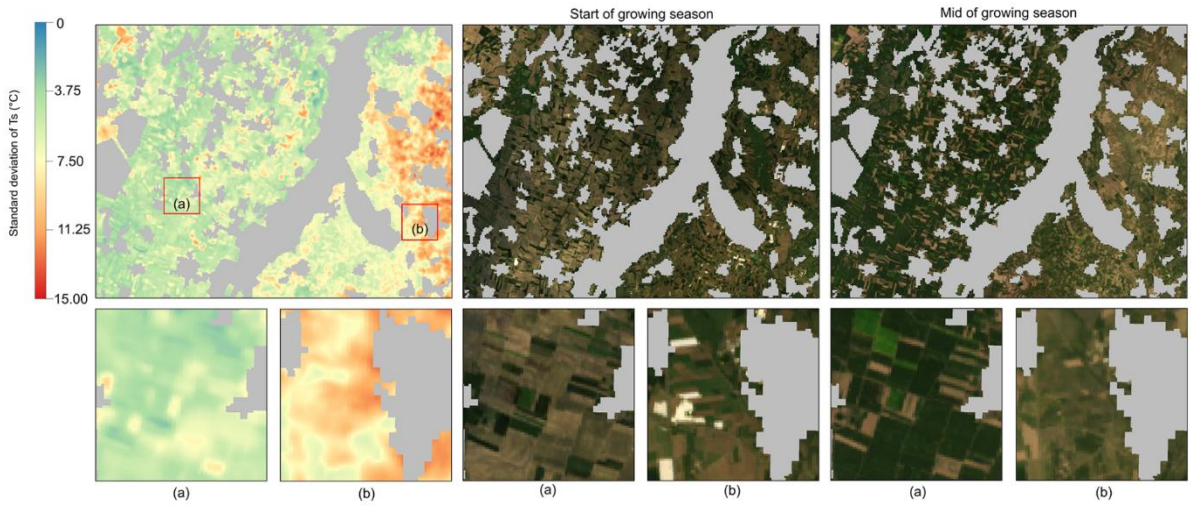


Figure 20. A snapshot of agricultural land shows the spatial distribution of standard deviation of land surface temperature throughout the growing season alongside true colour images (Landsat 7 and 8 imagery), highlighting (a) an area of irrigated cropland and (b) a non-irrigated area (Purnamasari et al., 2025).

To identify irrigated areas, the spatiotemporal features used for RF classification included the 10th percentile (p10), median (p50), 90th percentile (p90), mean, and standard deviation, calculated from the annual ΔT_s data cube. In this study, a pixel was identified as irrigated when it received at least one irrigation event. Thus, if the irrigation events were recurrent within a year, these events were counted as one. Pixels identified as irrigated only once throughout the study period were excluded, considering the high installation costs of irrigation equipment true irrigated pixels are unlikely to be irrigated only once in a nine-year period. The final classification results differentiate irrigated from non-irrigated pixels but does not estimate the fraction of irrigation within each pixel which may lead to overestimation or underestimation of irrigated areas.

Evaluation data

The mapped irrigated areas were evaluated against the total irrigated area obtained from the statistical office of the European Union, Eurostat, at NUTS level 2 for the year: 2013 and 2016 (data code in Eurostat website: ef_poirrig), available at https://doi.org/10.2908/EF_POIRRIG (last access: 20 June 2024). These statistics were collected from farm structure surveys (FSS). It is important to highlight that the methodologies and variables may vary across the EU member states, resulting in potential error in the validation data. The classification results were evaluated for overall, dry, wet NUTS2 regions which were defined based on the climatology of precipitation and potential evapotranspiration defined by (Purnamasari et al., 2025). The dry regions were classified as NUTS level 2 regions that lie within the Middle Rhine sub-basins. Meanwhile, the wet regions are in the Moselle, Neckar, Main, and Lower Rhine sub-basins (see Figure 18a). The estimated irrigated area and the reported area at NUTS level 2 were mapped to evaluate the accuracy of the model.

4.3 Results

Figure 21 shows an example time series of precipitation, evapotranspiration, simulated land surface temperature ($T_{s,sim}$), observed land surface temperature from MODIS ($T_{s,obs}$), and temperature difference (ΔT_s) for non-irrigated and non-irrigated pixels for training data collected from regions delineated in Figure 18. In Figure 21a, evapotranspiration gradually increases towards the peak of growing season in July, reaching higher magnitude than precipitation indicating a potential of water-limited regime. After reaching its peak, evapotranspiration gradually decreased toward the end of growing season. For irrigated pixel, the land surface temperature difference follows the pattern of evapotranspiration and peaks during growing season (Figure 21b). The complete absence of irrigation representation in the wflow_sbm model leads to higher land surface temperatures compared to observations. Assuming net radiation is the same for both, a greater portion of the net radiation is used for evapotranspiration (latent heat flux), while less is used for heating the air (sensible heat flux). On the contrary, for a pixel labelled as non-irrigated, simulated land surface temperature closely resembles observed land surface temperature which translates to similar partition of net radiation into sensible and latent heat flux (Figure 21c). Based on the temporal dynamics, we use simple statistical measures as features to identify irrigated pixels from non-irrigated area.

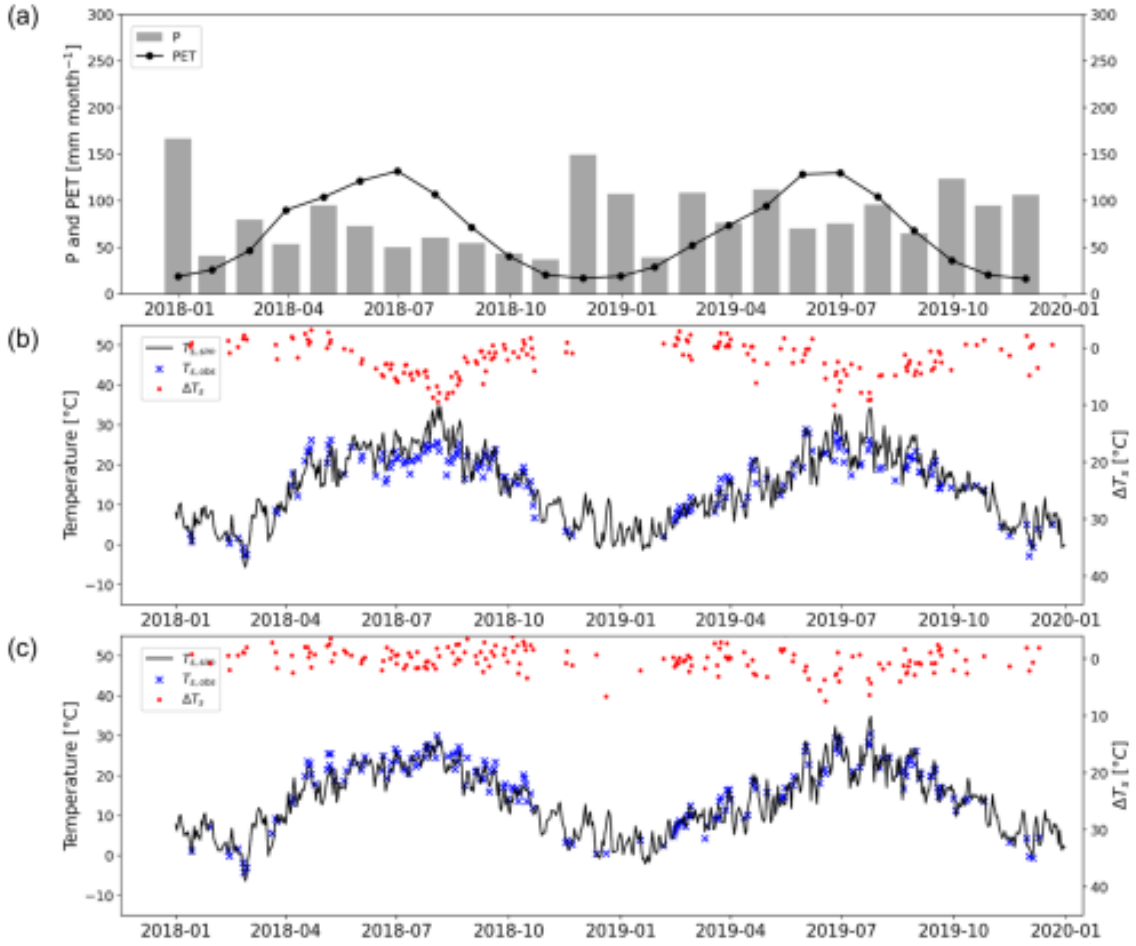


Figure 21. An example of time series of: (a) monthly precipitation (P) and evapotranspiration (PET) for the whole basin, and simulated land surface temperature ($T_{s,sim}$, black line), observed land surface temperature from MODIS ($T_{s,obs}$, blue crosses), and temperature difference (ΔT_s, red dots) for (b) an irrigated) and (c) a non-irrigated pixel derived from training data in 2018 to 2019 (Purnamasari et al., 2025).

Figure 22 shows the comparison between the mapped irrigated area estimated using the RF model and the reported irrigated area in Eurostat for 2013 (Figure 22a) and 2016 (Figure 22b). In general, the comparison shows a good agreement, with an R-squared values for the entire study area (R_{oa}^2) of 0.79 and 0.77 for 2013 and 2016, respectively. The R-squared values for the dry regions (R_{dr}^2) are higher for 2013 and 2016, thus indicating a better agreement. However, the R-squared value for wet regions (R_{wr}^2), are slightly lower. The estimated data consistently exceeds the reported data, with an average percentage relative difference of 17% compared to the subnational statistics. These overestimations were observed in areas with small agricultural holdings such as in Arnsberg (DEA5) and Koblenz (DEB1) (<10 hectares per holding). On the contrary, the methodology performs better in estimating irrigated area on regions with larger agricultural holdings (>22 hectares per agricultural holding) such as Düsseldorf (DEA1), Rheinhessen-Pfalz (DEB3), Köln (DEA2), and Darmstadt (DE71).

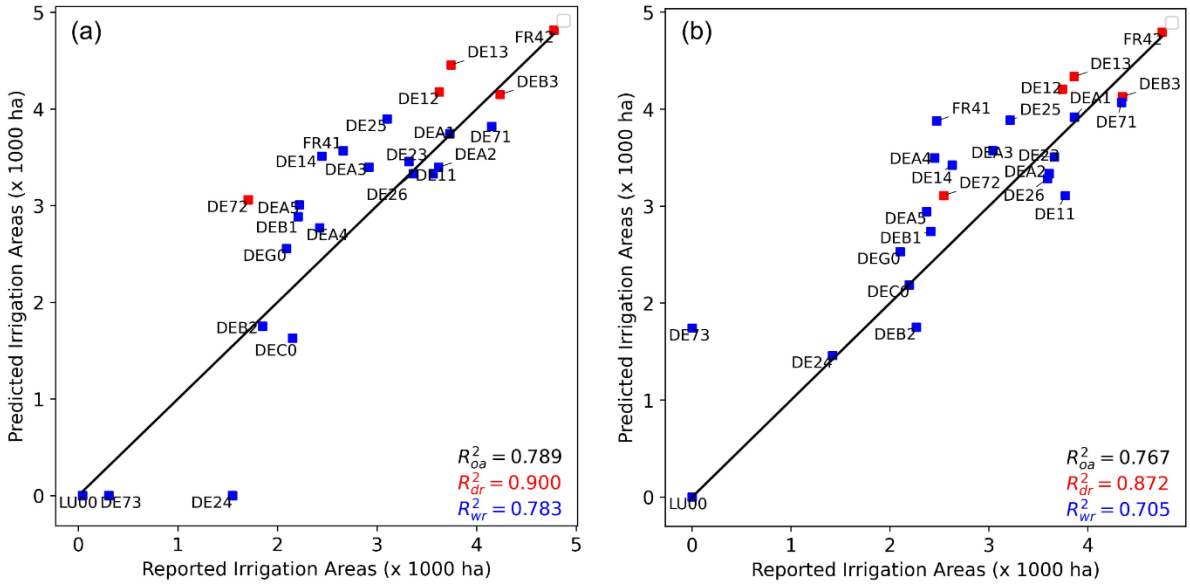


Figure 22. The estimated irrigated map compared with the total irrigated area reported in Eurostat data for the year: (a) 2013 and (b) 2016 (see Figure 18c for the Eurostat NUTS 2 regions). Area in hectares are transformed with log(y+1) transformation. R^2_{oa}, R^2_{dr} and R^2_{wr} denote the R-squared value for the entire study area (black), dry regions (red), and wet regions (blue) respectively (Purnamasari et al., 2025).

Figure 23 shows the irrigation frequency for the period 2010-2019 derived from the multiyear irrigated maps. As can be seen in Figure 23, irrigation hotspots were identified mainly in the Lower Rhine (b), the Middle Rhine (c), and the main subbasins of the Rhine (d). Based on the 10 annual maps, the total estimated irrigated area in the Rhine basin was found to be 170 thousand hectares (4.1% of the total area). From this portion, 1.5% of the area consistently received irrigation throughout the study period, and these pixels were mostly found in the Middle Rhine (Figure 23d).

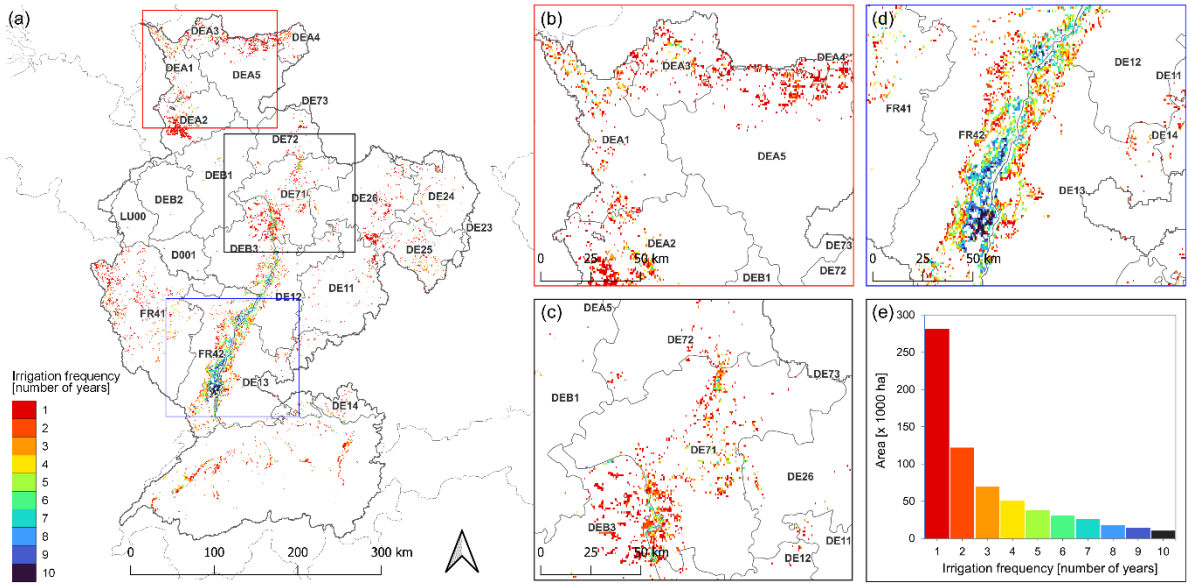


Figure 23. Irrigation frequency of pixels in the Rhine basin from the multiyear maps from 2010–2019, highlighting regions with the highest irrigation areas in the (a) Rhine basin, (b) Lower Rhine, (c) Middle Rhine, and (d) Rhine valley. Panel (e) shows the irrigation frequency and corresponding area (Purnamasari et al., 2025).

4.4 Discussion

The difference between our estimates and the irrigated area reported in subnational statistics can be attributed to the following factors: (a) a mismatch in spatial resolution between our simulations (1 km) and the subnational statistics of irrigated area and (b) uncertainties in the reported irrigated area. The pixels identified as either irrigated or non-irrigated are not adjusted based on the proportion corresponding to the size of agricultural holdings in a region. As a result, this may lead to either an underestimation or overestimation of the irrigated area in regions where agricultural holdings smaller than 1 km^2 are dominant. Meanwhile, the subnational statistics of irrigated areas were collected through questionnaires that were distributed to several agricultural holdings. Comparing continuous spatial information from classification results with point information obtained from questionnaires is not ideal for direct comparison. Additionally, there are uncertainties in the reported irrigated area due to sampling and non-sampling errors as the FSS involves random sampling methods and extrapolation techniques to produce data on irrigated areas.

4.5 Conclusion and next steps

This study uses an ML model to estimate irrigated areas in the Rhine basin on an annual timestep from 2010-2019, using land surface temperature from MODIS observations and a hydrological model to improve irrigated area estimates at a fine spatial scale (1 km). Although the results align well with subnational statistics, uncertainties remain due to the spatial resolution of the model. In regions where the actual irrigated areas are smaller than the spatial resolution of the model, this mismatch can lead to both overestimations and underestimations of irrigated area.

This work represents the first step in estimating agricultural water demand and will be used in further work to project future water needs under changing climate condition.

5 Predictive mapping of groundwater quality

5.1 Introduction

The goal of this research is to provide a predictive mapping tool for groundwater contamination, as improved modelling of groundwater resources and contamination was identified as a key topic for research by several of the RBHs (Duero and East Anglia) and may be useful to other basins. In a certain sense, the objective is similar to that of traditional geostatistics, except that ML allows users to account for many variables at the same time. The tool of choice for this research is MLMapper 2.0. This is a QGIS plugin developed by the team at the Universidad Complutense Madrid and subsequently improved in the context of this project with the collaboration of researchers at the UK Centre of Ecology and Hydrology (Martínez-Santos & Renard, 2019; Gómez-Escalonilla et al., 2022a). Mapping groundwater contamination in space serves various purposes. Results may be used to delineate the areas affected by contamination at the basin scale. These can also be useful in terms of explaining trends in groundwater contamination in different boreholes, as well as deciding where to locate new monitoring boreholes.

ML algorithms are inherently complex, and so are the relationships amongst explanatory variables that lead to a given target outcome. Thus, it is typically unfeasible to forecast whether one algorithm will outperform another, or whether an algorithm will yield good results on a dataset at all. To deal with this, MLMapper implements many supervised classification algorithms simultaneously, which then allows the modeller to select the best performing algorithm(s) for a given application. These all stem from the SciKit-Learn 0.24.1 toolbox (Pedregosa F. et al., 2011). MLMapper has proven suitable for predicting spatially distributed variables such as groundwater potential, borehole yield, nitrate pollution, and the presence of groundwater-dependent ecosystems (Martínez-Santos et al., 2021; Gómez-Escalonilla et al., 2022a; Gómez-Escalonilla et al., 2022b; Pacios et al., 2023).

For the purpose of this report, the method and its potential outcomes are illustrated through practical applications in the East Anglia and Duero RBHs.

5.2 Method

The goal of predictive mapping, MLMapper's main functionality, is to provide spatially-distributed predictions of a target variable based on a combination of predictor variables and a ground-truth database. This is done within a GIS environment. The underlying assumption is that a target variable, such as groundwater contamination, can be inferred from a series of indirect indicators (predictor variables) including lithology, land use, landforms, rainfall, lineaments, slope and drainage density, among others. A geographic database with ground-truth information can then be used as point-source input to train ML algorithms. Indeed, each borehole has a set of coordinates and a bivariate outcome, success or failure, as well as a "pixel-score" for each one of the predictor layers. If the borehole database presents a sufficiently large and diverse number of records, ML algorithms may be used to identify those patterns among the predictor variables that explain the presence or absence of contamination. The validated algorithm can then be extrapolated to obtain a groundwater contamination score for every pixel in the geographic database. This is, effectively, a predictive groundwater contamination map.

Figure 24 provides an overview of the predictive mapping routine implemented within MLMapper. Explanatory variables adopt the shape of layers in a GIS environment. Each algorithm attempts to find the combination of explanatory variable values in a pixel that leads to that pixel being classified as contaminated or non-contaminated (note that the threshold for contamination can vary for each application of MLMapper). Then, as every pixel value is known for every layer (predictor) in the GIS database, the combinations found by the algorithms can be extrapolated in space to develop a predictive map. Parameter weights are computed during this process to find out which explanatory variables can explain contamination more effectively. The software enables the user to pick one or several among a set of scoring metrics. These comprise raw test score, area under receiver operating characteristic curve (AUC), precision, recall, harmonic mean, and balanced score.

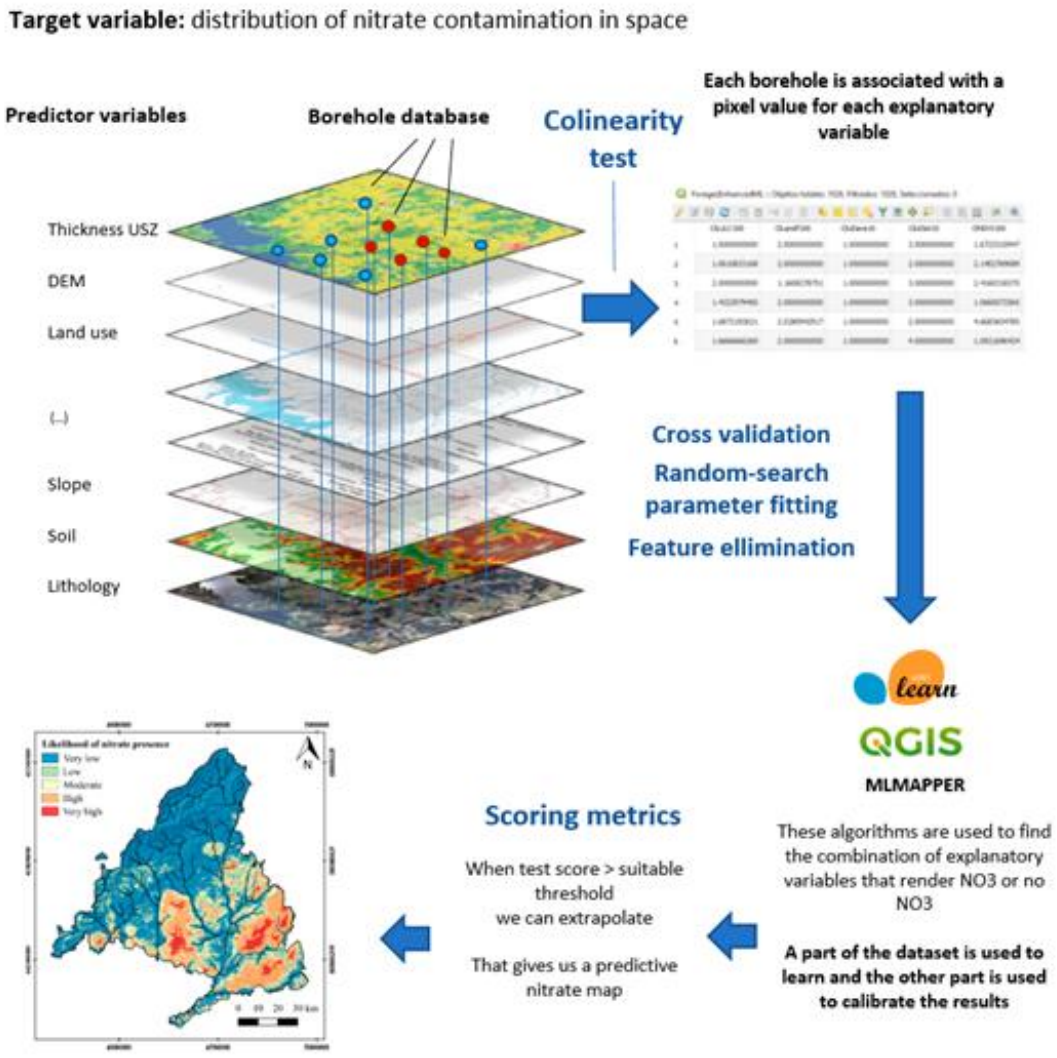


Figure 24. Simplified flowchart for MLMapper and its application to mapping nitrate (NO_3) contamination in groundwater.

Although MLMapper supports a wide range of classification algorithms, in this study we only used decision tree-based models. This choice is supported by a number of trial runs, as well as by previous studies demonstrating a better performance of this approach compared to other algorithms (Gómez-Escalonilla et al., 2024a; Gómez-Escalonilla et al., 2024b; Gómez-Escalonilla et al., 2024c). Additionally, decision trees offer advantages, such as not requiring data scaling and the ability to handle pixels with

missing values (i.e. no need to infill the missing data). Predictive maps were thus generated using the Gradient Boosting Classifier (GBC), Extra Trees Classifier (ETC), Random Forest Classifier (RFC), and AdaBoost Classifier (ABC) algorithms.

MLMapper selects the best performing algorithms based on user-defined performance thresholds. Algorithms performing poorly are discarded. The software routinely implements collinearity checks, randomized-search parameter fitting, cross validation and recursive feature elimination, to ensure a robust model. It also enables users to set the most suitable scoring metric in each case, as well as to carry out an ensemble of the best performing algorithms to estimate uncertainty.

Datasets

Borehole databases were made available by the Duero River Basin Authority and Anglian Water, respectively. Both include information on borehole coordinates, depth, wellscreens, GWL and nitrate content.

Nitrate data in water supply boreholes of the basin has been used as the target variable. To increase the representativity of the sample, we only used those boreholes with at least ten readings within the 2020-2024 period. MLMapper's architecture requires boreholes to be labelled as "contaminated" or "uncontaminated" (i.e. "positive" or "negative" class, respectively). Therefore, a contamination threshold was determined for each case based on environmental and drinking water regulations. This amounts to 37.5 mg/l for the Duero basin and 50 mg/l for the Anglian region. As nitrate concentration in groundwater tends to oscillate over time, and also because data series span several years in both cases, certain boreholes were observed to exceed the threshold either occasionally or frequently. In these cases, boreholes were considered contaminated whenever nitrate content exceeded the threshold at least 10% of the time. This percentile was determined in agreement with the stakeholders.

While the goal is the same for the two areas, the hydrogeological setting and the practical purpose is slightly different in each case, which constrains the number of available data points. In the Duero basin we attempt to map nitrate content for the whole geographical domain. This includes different hydrogeological conditions and aquifer types, including confined, leaky, and unconfined units, as well as fissured, karstic, and detrital aquifers. Thus, the Duero database included 284 points distributed across the whole basin. Approximately 67% of these were contaminated. In contrast, the aim in the East Anglian region was to map the presence of nitrate only in the chalk aquifer systems. These are relatively homogeneous from the hydrogeological point of view, but still feature unconfined conditions towards the eastern border and confined conditions in the central and western parts. The Anglian database included 301 points, out of which 39% were labelled as contaminated. All boreholes in non-chalk areas of the region were discarded.

5.3 Results

Duero basin

The Duero model was trained with 284 points prior to extrapolating the results to the whole of the study area (Figure 25). Table 9 presents the probability of groundwater exceeding 37.5 mg/l of nitrate as per the four tree-based algorithms. It includes the accuracy values for each algorithm, an indicator that reflects the proportion of correct predictions made by the model out of the total number of

predictions. All models achieved high accuracy on the training set, with values between 91 and 100%, which suggests some degree of overfitting, particularly in the case of GBC and ETC. On the test set, accuracy values ranged between 86% and 90%, indicating good performance in terms of model generalization.

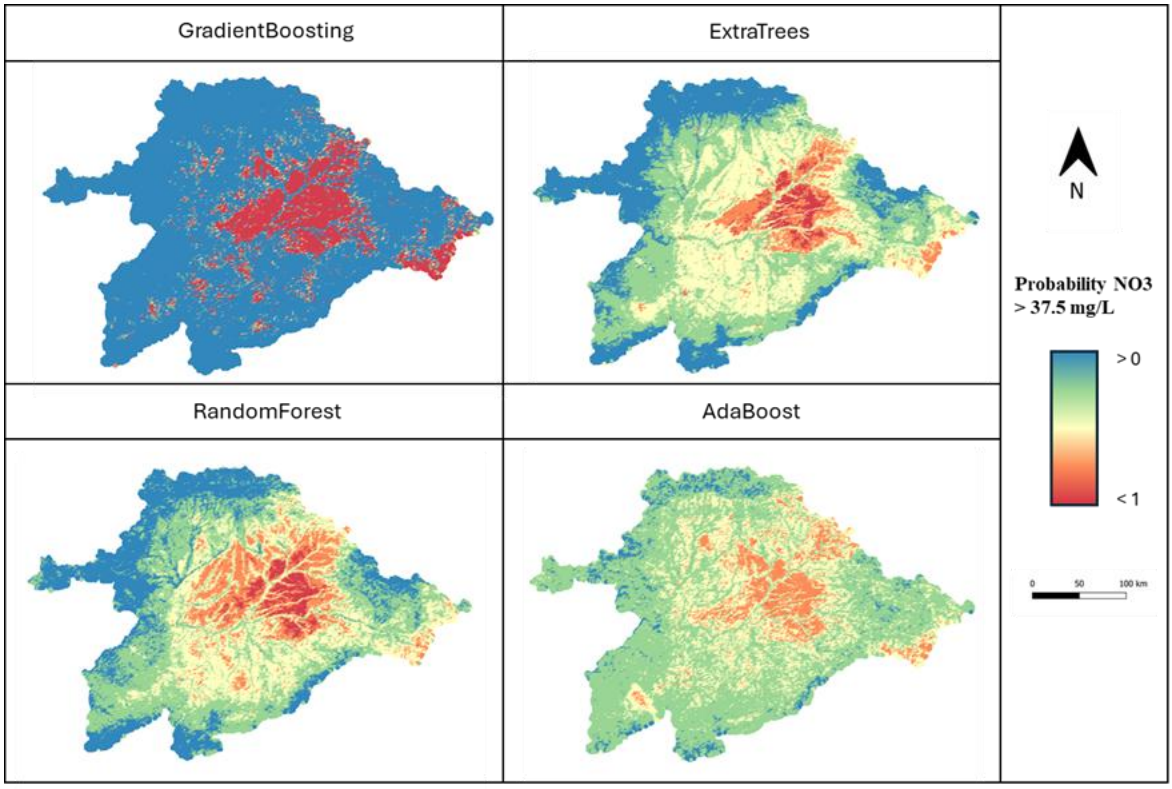


Figure 25. Predictive groundwater vulnerability map for the Duero basin. Pixel-scale values are expressed as the probability of exceeding 37.5 mg/l of nitrate (NO_3).

Table 9. Summary of machine learning metrics in the Duero basin for the four algorithms, namely Gradient Boosting Classifier (GBC), Extra Trees Classifier (ETC), Random Forest Classifier (RFC), and AdaBoost Classifier (ABC) . Performance metrics are: Train score = optimised training score; Test score = optimised test score; Prec 0 = precision negative class; Prec 1 = precision positive class; 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	F1 0	F1 1	AUC
GBC	1.00	0.86	0.87	0.67	0.89	0.62	0.79
ETC	1.00	0.90	0.91	0.83	0.93	0.76	0.95
RFC	0.91	0.90	0.95	0.75	0.93	0.80	0.90
ABC	0.83	0.86	0.95	0.67	0.90	0.75	0.94

Within the ground truth dataset, there is a noticeable imbalance between the negative class (66%), samples with nitrate concentrations below 37.5 mg/l and the positive class (34%), samples with nitrate concentration above 37.5 mg/l. It is therefore necessary to analyse each class separately.

The F1 scores were higher for the negative class (ranging from 0.89 to 0.93) than for the positive class (ranging from 0.62 to 0.80), indicating better detection of non-contaminated cases. Both RFC and ETC outperformed GBC and ABC across these metrics. The Area Under the Receiver Operating Characteristic curve (AUC) is a single value that summarizes the performance of a binary classifier. It

represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. AUC values range from 0.5 to 1.0, where 0.5 indicates random guessing and 1.0 indicates perfect classification. In this case, results ranged from 0.79 for GBC, to 0.94 for ABC.

The RFC model was selected to apply recursive feature elimination to identify which explanatory variables contribute most significantly to predictions. Figure 26 suggests that RFC relies mostly on variables such as sand/clay content, distance to permanent surface water (alluvial sediments), NDVI, and distance to irrigated plots. In the case of sand and clay content, this can be explained by the fact that the central areas of the basin (detrital aquifers) are comparatively more contaminated than the remainder of the study area. The exception to this rule is the alluvial sediments that occur along permanent water courses. This is also consistent with variable importances and can be attributed to the fact that these sediments are highly permeable, which means that contamination is flushed off quickly every season. The correlation between contamination and predictors such as landforms, slope, and land use, is typically low.

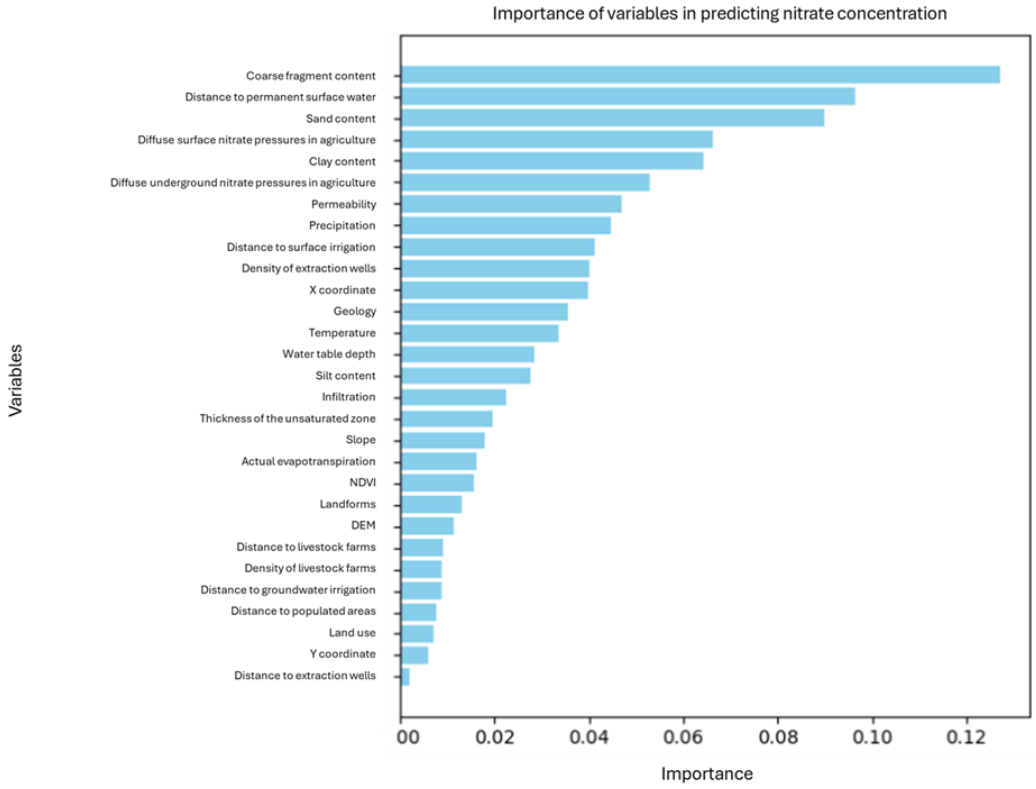


Figure 26. (Normalised) Feature importance in the Random Forest classifier obtained by applying Recursive Feature Elimination.

East Anglia basin

In the case of East Anglia, groundwater contamination was mapped based on explanatory variables such as land use, landforms, lithology, and precipitation. Tree-based algorithms outperformed other algorithm families. The test scores and area under the receiver operator characteristic curve scores for RF and ET are around 0.90.

The predictive performance of the classification algorithms is shown in Table 10, which lists the main metrics obtained by the ensemble tree-based algorithms applied. The outcomes indicate a high predictive capability of the four models. The test scores all exceed 0.83, with three of the algorithms (RFC, ABC and GBC) achieving a score above 0.89, indicating a significant predictive accuracy. The analysis of the F1-score indicates that the algorithms are effective in predicting both classes, as evidenced by the F1-score values for the 1 (contaminated) and 0 (not contaminated) categories that consistently exceed 0.8, with a maximum value of 0.96 observed for GBC. The AUC values, ranging from 0.96 to 1 for all models, demonstrate a high predictive performance, as these values represent the upper end of the spectrum of predictive performance with classification models.

Table 10. Summary of machine learning metrics for the tree-based ensemble models in the East Anglian chalk aquifer. Performance metrics: Train score = optimised training score; Test score = optimised test score; Prec 0 = precision negative class; Prec 1 = precision positive class; Rec 0 = recall negative class; Rec 1 = recall positive class; 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

Figure 27 illustrates the mean probability predicted by the four algorithms to exceed the 50 mg/L threshold. The highest probabilities are found in the western part of the chalk aquifer extension, where these materials outcrop at the surface. In the central areas, the aquifer materials are confined by quaternary materials, which provide some protection from sources of surface nitrate contamination. In the eastern band, streams erode quaternary materials and allow chalk to reach the surface again, increasing the likelihood of exceeding the nitrate contamination threshold. In general, most contaminated and uncontaminated points can be correctly identified by the algorithms, even in regions where these points are near each other. This is the case of two points (one contaminated and one uncontaminated) shown in Figure 27, where we can see how the algorithms offer a higher probability around the contaminated point than in the area corresponding to the other borehole.

Figure 28 provides a more comprehensive understanding of the factors that, according to the ET algorithms, control nitrate pollution in the test boreholes. Figure 28a presents a bee swarm plot obtained with the SHapley Additive exPlanations (SHAP) technique (Nohara et al., 2022). SHAP analyses facilitate the identification of the most important variables (from top to bottom in order of importance), as well as the identification of the values of these variables that have a positive impact (greater probability of contamination) or a negative impact on the prediction of nitrate content.

In this case, buried valleys, the thickness of quaternary materials and the elevation of the base of the chalk aquifer rank among the most important variables. In addition, it highlights how higher values of the first two variables are related to a negative impact on the predictions (i.e., lower probability of contamination) while higher values of the base of the aquifer result in a positive impact (higher probability of contamination) on the predictions. Figure 28(b) and (c) present the SHAP values for a couple of individual pixels, those identified in Figure 27. These explain how the algorithms rely on different variables to explain why two boreholes that are relatively close in space may present entirely different outcomes in terms of nitrate contamination.

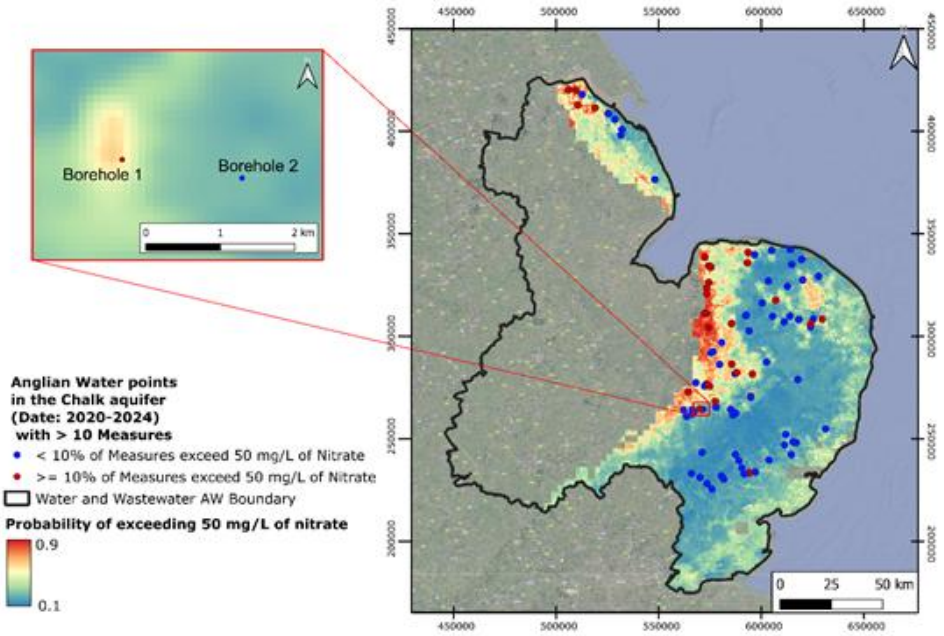


Figure 27. Predicted groundwater contamination from nitrate in the chalk aquifers of East Anglia.

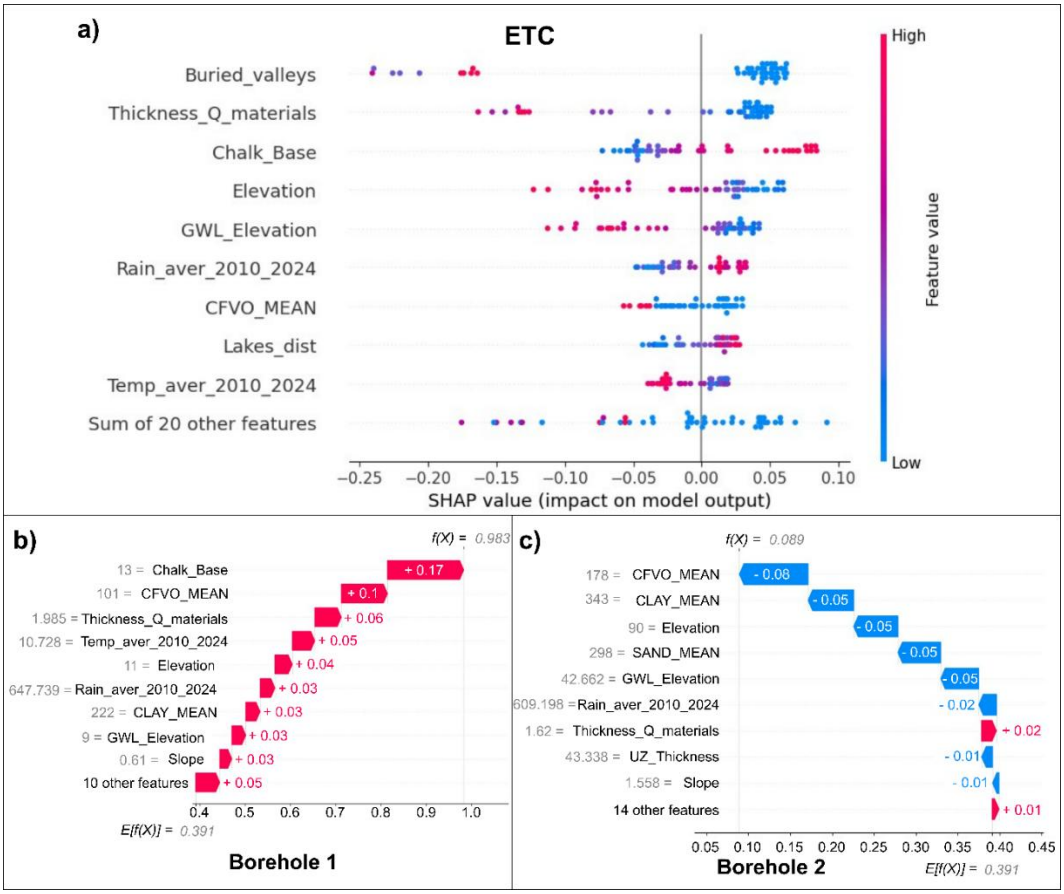


Figure 28. (a) Beeswarm plot of the SHAP Analysis obtained for the Extra Trees Classifier (ETC) model for the entire test data set. (b) Waterfall plot of the SHAP Analysis for the Borehole1 and (c) Borehole 2, borehole locations can be seen on Figure 27.

5.4 Discussion

In the case of the Duero basin, there is a strong correlation between predictive maps and prior knowledge. Nitrate contamination is more prevalent in the northeastern quadrant, an area characterized by the presence of permeable sediments and widespread agriculture. Conversely, basin boundaries, made up mostly of impervious materials, are predicted to be contamination-free. Alluvial systems are clearly identifiable, particularly in the case of the RF and ET algorithms. These show up as less contaminated than the surrounding areas, which is also consistent with field experience, and which could be attributed to the fact that alluvial systems are typically more permeable. This implies that contamination washes off more quickly than in the interfluve areas.

This interpretation is consistent with the feature importance analysis. Key variables to explain nitrate contamination in the Duero basin include sediment types (sand, clay, and, to a lesser extent, silt), and distance to permanent surface water. Agriculture-related variables such as normalized difference vegetation index (NDVI) and distance to agricultural areas also rank among the most important predictors for groundwater contamination. In contrast, digital elevation model (DEM)-derived variables such as topography, slope, and landforms were found to be relatively unimportant. This appears to be because the areas characterized by the presence of permeable materials are mostly flat, which means that topography-related predictors provide relatively little valuable information to the algorithms.

In the case of East Anglia, the resulting map is consistent with both the conceptual groundwater model and field information. In particular, it adequately depicts the existence of nitrate-vulnerable areas along chalk aquifer outcrops. There is a marked contrast between these areas and those where the productive aquifers are protected by overlying impervious or semi-impervious materials. Furthermore, the map adequately identifies deeply incised valleys towards the eastern side of the basin, where the conditions are similar to the chalk outcrops.

5.5 Conclusions and next steps

This assessment confirms the potential of ML tools to enhance the spatial extent of groundwater contamination predictions across different hydrogeological settings. While the outcomes can still be further refined, results suggest that the algorithms are already able to depict field conditions reliably. The fact that each case relies on different predictors, together with the fact that the hydrogeological settings are significantly different, suggests that this method can be readily exported to any settings provided there is enough ground truth data available for training and testing the algorithms. The method caters to a wide variety of field conditions as the choice of predictor variables can be modified by the user to account for site-specific considerations. Future developments potentially include work with contaminants other than nitrate. This is expected to enhance our understanding of groundwater contamination problems across different basins.

Because ML algorithms learn from experience, predictions must necessarily be based on a ground-truth dataset. In this context, it should be noted that having a pre-existing borehole network is essential to apply MLMapper.

6 Quantitative groundwater resources estimation

6.1 Introduction

Regarding the quantitative estimation of groundwater resources, significant challenges face the Duero basin, challenges that may also resonate in basins like Messara and Seine. These challenges include:

- Adapting to dynamic scenarios: Navigating through evolving scenarios involving new agricultural practices, changing demands, increased solar irrigation, and the effects of climate change.
- Enhancing resource utilization efficiency: Improving the efficiency of resource utilization to meet growing demands sustainably.
- Implementing real-time groundwater monitoring: Establishing a robust system for real-time monitoring of GWLs at a comprehensive scale.

Consequently, there is a recognized imperative to refine the estimation of groundwater resources, as acknowledged by stakeholders in the Duero. This refinement should occur at a higher resolution than what current process-based models offer, encompassing both spatial considerations (e.g., sub-basin scale, groundwater basin scale) and temporal dimensions (e.g., weekly, biweekly, monthly).

However, the existing process-based hydrogeological models encounter limitations in achieving these objectives. Challenges include the approaching computational thresholds for higher spatial resolution and larger ensemble scenario predictions. Additionally, these models often underperform in regions with limited observational data for calibration, failing to adequately capture the spatial dependence structure of hydrologic processes across various spatiotemporal scales.

In response to these challenges, there is a concerted effort to explore the potential of a data-driven surrogate meta-model. This type of model aims to predict across diverse spatial resolutions. Similar methodologies have shown promise in predicting streamflow, e.g. (Arsenault et al., 2023), and water balance components, e.g. (Droppers et al., 2023).

To our knowledge, the only related previous work using this type of model in the context of groundwater is by Seo and Lee (2021). In this study, they tested the application of LSTM and Convolutional LSTM (ConvLSTM) deep learning (DL) architectures to predict changes in spatiotemporal GWS in South Korea. To achieve this, they used gridded precipitation, temperature, soil moisture, NDVI and GRACE-derived TW) as inputs, but at a relatively coarse resolution (0.25°, 1 month). They observed that the ConvLSTM model captured spatiotemporal variations in GWS better than the LSTM model.

Therefore, applying such approaches to groundwater balance components at high spatiotemporal resolution remains a formidable challenge. Our objective is to develop a robust DL surrogate model capable of reproducing spatiotemporal variations in GWS in the Duero Basin, which serves as a pilot site. Furthermore, the model(s) should ideally be transferable to other basins and adaptable to their respective data and local conditions.

6.2 Method

Figure 29 illustrates the workflow for developing DL models to estimate groundwater storage Change (GWSC). The process begins by collecting static and dynamic spatiotemporal data layers that are directly or indirectly related to groundwater dynamics from both local and global sources (as detailed in Table 11). To ensure compatibility with the DL models, these input layers and the target GWSC data are spatially and temporally resampled and rescaled to a unified resolution. The workflow also incorporates a feedback loop for continuous improvement, which allows for the iterative re-selection and reprocessing of variables.

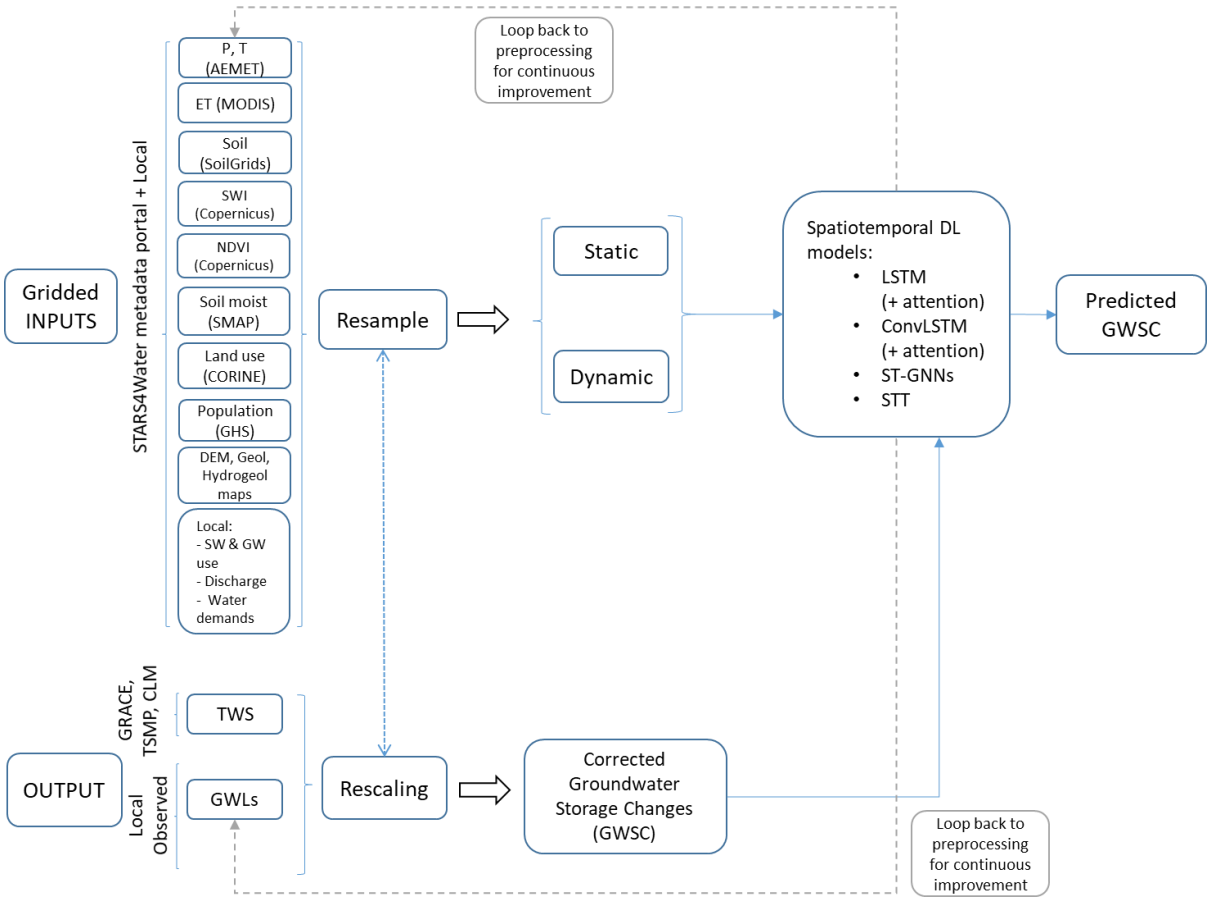


Figure 29. Workflow for Deep Learning modelling of Groundwater Storage Changes (GWSC) in the Duero Basin.

Based on current team knowledge and the literature reviewed, the envisaged surrogate model is a DL variant like LSTM or Convolutional LSTM (ConvLSTM) with attention (Figure 29). LSTM is a type of recurrent neural network specifically designed to capture long-range temporal dependencies in sequential data. ConvLSTM extends LSTM by replacing the fully connected operations with convolutional operations in the input-to-state and state-to-state transitions. This allows the model to capture both spatial and temporal dependencies, making it well-suited for spatiotemporal data. Attention mechanisms enhance LSTM or ConvLSTM by allowing the model to dynamically focus on the most relevant parts of the input sequence (or spatial-temporal features) during prediction. This improves model interpretability and performance, particularly in complex tasks involving heterogeneous input data.

On the other hand, novel architectures such as Spatiotemporal Graph Neural Networks (ST-GNN) and Spatiotemporal Transformer (STT) were also tested. ST-GNNs are DL models designed to capture both spatial and temporal dependencies in data structured as graphs. In these networks, nodes (e.g., locations or sensors) are connected based on spatial relationships, and graph convolutions are applied to extract spatial features. Temporal dynamics are modelled through recurrent units (like LSTM) or temporal convolutions. STTs are a class of Transformer-based models that capture spatiotemporal dependencies using self-attention mechanisms. Unlike recurrent neural networks or GNNs, transformers model global interactions across both space and time without relying on sequential processing. STTs can efficiently learn long-range dependencies and are particularly powerful when applied to gridded or multivariate spatiotemporal datasets. Their flexibility and scalability make them promising candidates for complex geoscientific tasks such as GWSC prediction.

Shallow ML algorithms such as RF and Extreme Gradient Boosting (XGBoost) were also tested as benchmarks and to build ensembles with the DL methods.

The model is designed to forecast GWSC at a specific spatiotemporal scale and for a predefined forecast period. Predictions rely on input variables such as streamflow, GWL, and discharge, in tandem with various gridded static and dynamic spatiotemporal inputs accessible within the STARS4Water metadata portal (https://stars4water.eu/stars4water-metadata-portal/). These inputs encompass, amongst others, static geological, geographic, hydrological, hydrogeological and socio-economic characteristics, such as soil composition, precipitation, evapotranspiration, soil moisture, NDVI, population density, water demands, and land use (Table 11).

Presently, the Duero basin managers use groundwater balance outputs derived from a process-based model (Patrical, https://iiama.webs.upv.es/en/technology-transfer/software/patrical/). This data should serve as valuable training material for the proposed DL model. However, the simulated data is not free of error and uncertainty, as inferred from the discrepancies between observed and simulated discharge flow rates, likely resulting from coarse scales and hydrogeological assumptions.

Given the considerable uncertainties associated with the Patrical model, its GWS simulations were deemed unsuitable as training data for the DL model. Instead, GWS estimates were derived by integrating water storage outputs from land surface models (LSMs) and hydrological models (HMs) with in situ GWL observations collected across the Duero Basin by the champion stakeholder, the Duero River Basin Authority. These water storage outputs include TWS, which is the sum of all water storage components considered in the model (e.g. groundwater, soil moisture, surface water, canopy water and snow water). A non-parametric Spearman correlation analysis was performed to assess the relationship between local GWLs and water storage outputs from various LSMs/HMs, as well as GRACE satellite data. Among the models evaluated, the Community Land Model (CLM) and the Terrestrial Systems Modelling Platform (TSMP) exhibited the strongest correlations with GWLs (WTDA_CHD) at the coarsest spatial resolutions (Figure 30). Note that WTDA are negatively correlated to Total Water Storage Anomalies (TWSA); Liquid Water Equivalent Anomalies (LWEA); and Runoff Water Equivalent (ROWE), since an increase in water table depth means a decrease in stored water.

Table 11. Groups of layer variables used to train the Deep Learning model for Groundwater Storage Change (GWSC) estimation in the Duero basin.

Layer Type	Properties	Nº Vars	Original Spatial Resolution	Original Temporal Resolution	Source	Reference Source
Geographic	Static	2	200 m	N/A	Local	CNIG
Socio-economic	Static	8	Vector (Corine: 100m)	N/A	Local (Corine: global)	Mirame Duero (Corine: land Copernicus)
Hydrology	Static	4	Vector	N/A	Local	Mirame duero
Hydrogeology	Static	6	Vector (Perm: 200m)	N/A	Local	Mirame Duero (Perm: IGME)
Soil	Static	2	Vector/250m	N/A	Local/global	Mirame Duero/soil grids
Geology	Static	1	50 m	N/A	Local	IGME
Vegetation	Static	1	Vector	N/A	Local	Mirame Duero
GHS Density POP	Dynamic	1	30 arcsec (~600 m)	Average every 5 years 2000-2025	Global	https://data.jrc.ec.europa.eu/dataset/2ff68a52-5b5b-4a22-8f40-c41da8332cfe
ET (MODIS)	Dynamic	2	500 m	Every 8 days 2000-2023	Global	https://human-settlement.emergency.copernicus.eu/download.php?ds=pop
NDVI Copernicus	Dynamic	1	1 km	Every 10 days 2015-2020	Global	https://lpdaac.usgs.gov/products/mod16a2gfv061/
SWI Copernicus	Dynamic	4	1 km	Daily 2015-2024	Global	https://land.copernicus.eu/en/products/vegetation/normalised-difference-vegetation-index-v3-0-1km
SMAP (Surface and Root Zone Soil Moisture)	Dynamic	21	9 km	Every 3 hours 2015-2024	Global	https://land.copernicus.eu/en/products/soil-moisture/daily-soil-water-index-europe-v1-0-1km
Agricultural and livestock demand	Dynamic	2	Vector	Monthly Jan.2014 to Dec. 2021	Local	https://nsidc.org/data/spl4smgp/versions/7
Supply demands	Dynamic	1	Vector	Monthly Jan.2013 to Dec. 2022	Local	Mirame Duero
GWL observations	Dynamic	1	point	Monthly	Local	https://www.chduero.es/red-de-control-del-nivel
Corrected TWS CLM	Dynamic	3	3 km	Monthly Dec. 2013 to Dec. 2018	Global	https://datapub.fz-juelich.de/slts/4DHydro/
Corrected TWS TSMP	Dynamic	1	11 km	Monthly	Global	https://datapub.fz-juelich.de/slts/4DHydro/

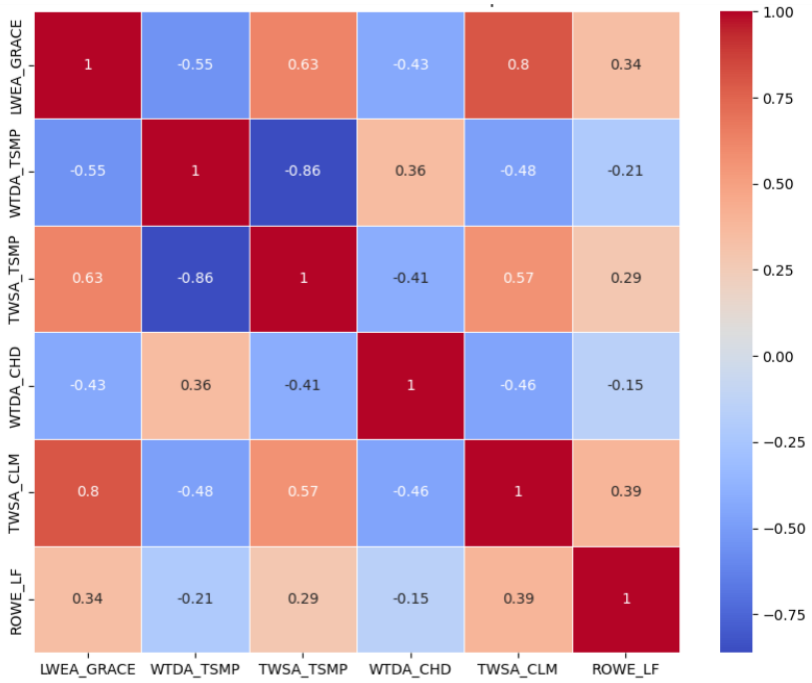


Figure 30. Spearman correlation matrix between water storage anomalies and water table depth anomalies (WTDA) data from different LSM/HMs and local observations (WTDA_CHD). TWSA: total water storage anomalies; LWEA: liquid water equivalent anomalies; ROWE: runoff water equivalent; LF: LISFLOOD.

Based on the correlation analysis, simulated TWS data from TSMP and CLM were integrated and corrected using local GWL observations from the Duero River Basin monitoring network at a monthly timestep as a data assimilation approach to develop GWS target variables. TSMP offers long-term (1989-current) monthly simulations of WTD and TWS across Europe at approximately 11 km resolution. Moreover, daily TWS from the CLM at 3 km resolution is also available at Pan-European level. The integration of CLM and TSMP data was performed using geostatistical interpolation and linear regression, as detailed below.

Correction of Community Land Model (CLM) Total Water Storage (TWS)

Simulated TWS from CLM exhibited an overall Spearman correlation of -0.46 with observed WTD across the entire Duero Basin (Table 9). The correction of TWS values from CLM was based on the linear relationship between TWS and WTD. Since multiple observations could fall within a single 3 km TWS pixel, the observation well with the highest correlation with TWS in each pixel was selected, and a linear regression model was fitted between TWS and observed WTD.

The regression equation was then used to estimate and correct TWS from WTD data at the pixel level, while an uncertainty flag variable was assigned based on the correlation coefficient (r) as follows:

- Low uncertainty: If $|r| > 0.7$, corrected values were computed using the regression line.
- Moderate uncertainty: If $0.5 < |r| < 0.7$, corrected values were computed using the regression line.
- High uncertainty: If $|r| < 0.5$ or no observation well was present in the pixel, no correction was applied to the CLM TWS values, and they were flagged as having "high" uncertainty.

Correction of Terrestrial Systems Modelling Platform (TSMP) Total Water Storage (TWS)

Apart from TWS, the TSMP model also provides simulated WTD data. TWS and WTD from TSMP exhibited an overall Spearman correlation of -0.41 and 0.36 with observed WTD across the entire Duero Basin, respectively (Table 9).

A more advanced correction approach was applied to the TWS data from TSMP, incorporating the interpolation of WTD residuals using spatiotemporal kriging. These residuals were then added to the average simulated WTD from TSMP (Ben-Salem et al., 2023). This interpolation method enhances the alignment of TSMP model outputs with observed WTD data.

The conditioning process is expressed as follows:

$$WTD_{TSMP(corrected)} = WTD_{TSMP} + \sum_{i=1}^{N} w_i [WTD_{o,i} - WTD_{TSMP,i}]$$

where $WTD_{TSMP(corrected)}$ is the corrected WTD from the TSMP model at each TSMP pixel, WTD_{TSMP} are the simulated WTD values from TSMP, $WTD_{o,i}$ - $WTD_{TSMP,i}$ are the residuals between observed and simulated WTD at and in a neighbourhood of the TSMP pixel, w_i represents the weights assigned during the kriging interpolation process, and N is the number of nearby observations used in conditioning.

After computing residuals between observed and simulated WTD values, the dataset was assigned to a continuous spatial grid by creating spatial bins based on the original X_CHD and Y_CHD coordinates. For each bin, the median values were computed to obtain a representative sample across the domain, ensuring a uniform spatial distribution. Next, grid cells were reassigned using the centres defined by X_TSMP and Y_TSMP, ensuring that every data point was mapped to a continuous spatial grid matching the 11 km spatial resolution of TSMP.

A crucial aspect of this analysis is its explicit incorporation of time. An experimental variogram was estimated from the representative data, and several variogram models were evaluated, using the Python package GSTools (Müller et al., 2022). The model that yielded the highest R^2 score between the experimental variogram and the model-predicted variogram was selected as the global model.

We then performed local kriging for each grid cell using this global model. For each cell, local data is selected based on spatial proximity, and ordinary kriging is applied to interpolate the residuals over time. The predicted residuals are added to the modelled values (WTD_TSMP) to produce a corrected time series.

Finally, corrected GWS from TSMP were derived at each pixel where corrected groundwater levels were available:

$$GWSC_{TSMP(corrected)} = S_j \cdot \Delta WTD_{TSMP\ (corrected)}$$

where $GWSC_{TSMP(corrected)}$ are the corrected values of GWSC for TSMP simulations, S_j are storage coefficients at each pixel j, estimated as the slope of the linear regression between the time series of ΔTWS and ΔWTD simulated by TSMP*, and* $WTD_{TSMP(corrected)}$ are the corrected WTD values for TSMP obtained from the kriging through the previous equation.

Similarly to the CLM case, uncertainty of GWSC corrections from TSMP simulations has been incorporated at the pixel level according to the following criteria:

- Low uncertainty: If ($R^2 > 0.7$ & $S < 0.3$) & $GWSC_{TSMP(corrected)}$ estimate available, $GWSC_{TSMP(corrected)}$ assigned as target value.
- Moderate uncertainty: If ($R^2 < 0.7$ or $S > 0.3$) & $WTD_{TSMP(corrected)}$ estimate available, simulated ΔTWS_{TSMP} assigned as target value.
- High uncertainty: If ($R^2 < 0.7$ or $S > 0.3$) & $WTD_{TSMP(corrected)}$ estimate not available, simulated ΔTWS_{TSMP} assigned as target value.

6.3 Results

Correction of CLM Total Water Storage

Figure 31 shows two examples of linear regressions between observed WTD in piezometers and simulated TWS from CLM in two groundwater bodies of the Duero basin. The strength of these linear regressions as measured by the correlation coefficients were then used for TWS data correction and uncertainty flagging, as explained in the Methods section.

Figure 32 shows the resulting uncertainty map for the TWS from CLM across the whole Duero River basin and an example of corrected TWS values for December 2013. The large area of high uncertainty in the left map is conditioned by the large amount of CLM pixels with no intersecting observation wells.

Correction of TSMP Total Water Storage

The experimental semi-variogram of the residuals between observed WTD in the monitoring network and the simulated WTD from the TSMP model is shown in Figure 33. The stable covariance model (Wackernagel, 2003) presents the highest goodness-of-fit (R2=0.93) to estimate the spatial correlations of the residuals at different lag distances.

This covariance model is then used as the global model to perform spatiotemporal ordinary kriging of the residuals at each TSMP grid cell (11 km), and the interpolated residuals are subsequently added to the simulated WTD from TSMP to generate the corrected time series of WTD values. A sample plot from one cell displaying the observed (WTD_CHD), modelled (WTD_TSMP), and corrected GWLs over time is shown in Figure 34. Finally, GWSC was estimated and uncertainty flagged at the pixel level, following the procedure described in the Methods section (Figure 29). There highest concentration of TSMP pixels with low uncertainty GWSC estimations are concentrated in the southern and central areas of the Duero basin.

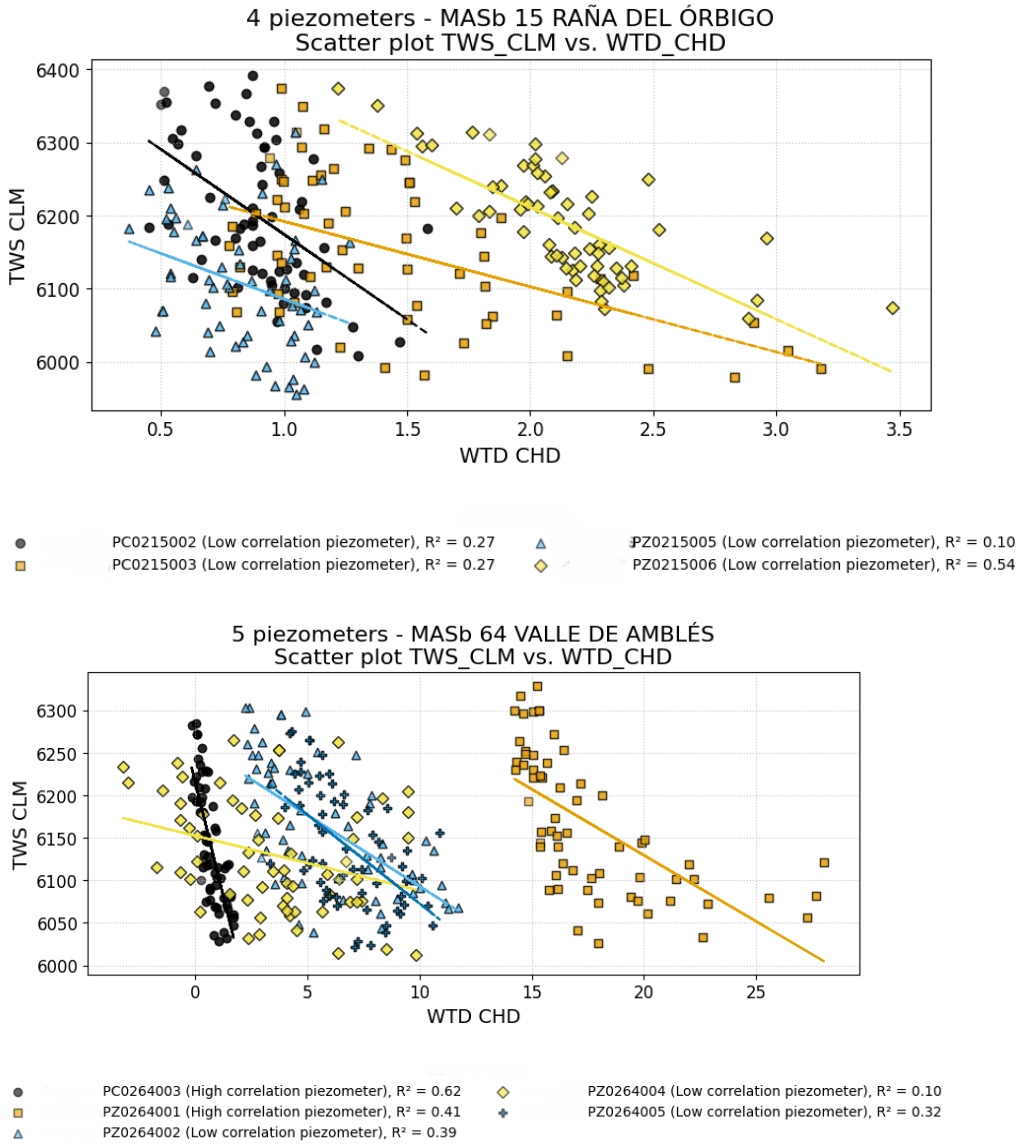


Figure 31. Linear regression lines between total water storage (TWS) from the CLM model, in mm, and local water table depth observations (WTD_CHD), in m, for different observation wells (piezometers) in (a) a shallow groundwater body, namely Raña del Orbigo, and (b) a deep groundwater body, namely Valle de Amblès, within the Duero River Basin.

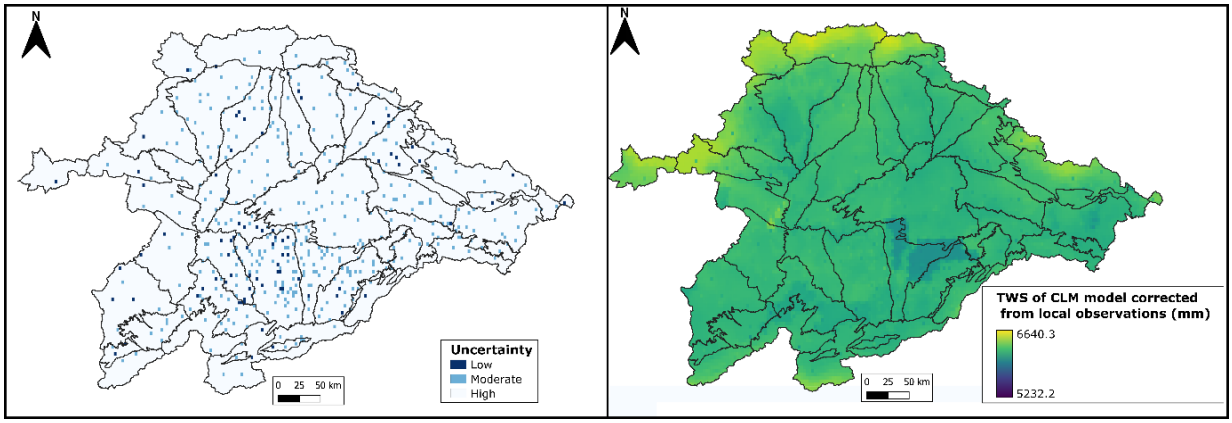


Figure 32. Left: Uncertainty map based on the correlation between observed WTD and TWS values from CLM. Right: Corrected TWS values from CLM for December 2013, derived using local WTD observations following the approach described in the Method section.

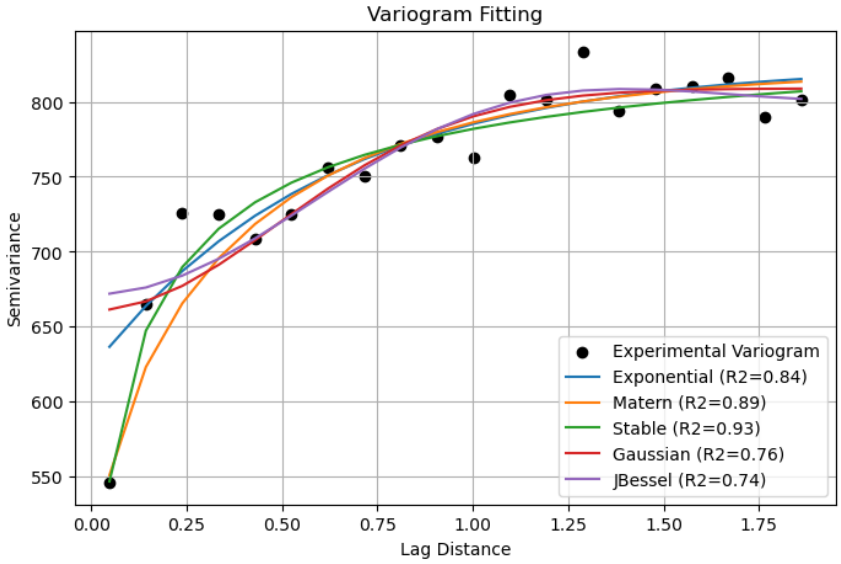


Figure 33. Experimental variogram alongside fitted covariance models with their R^2 values, for spatiotemporal kriging of the residuals between observed (WTD_CHD) and simulated (WTD_TSMP) water table depth.

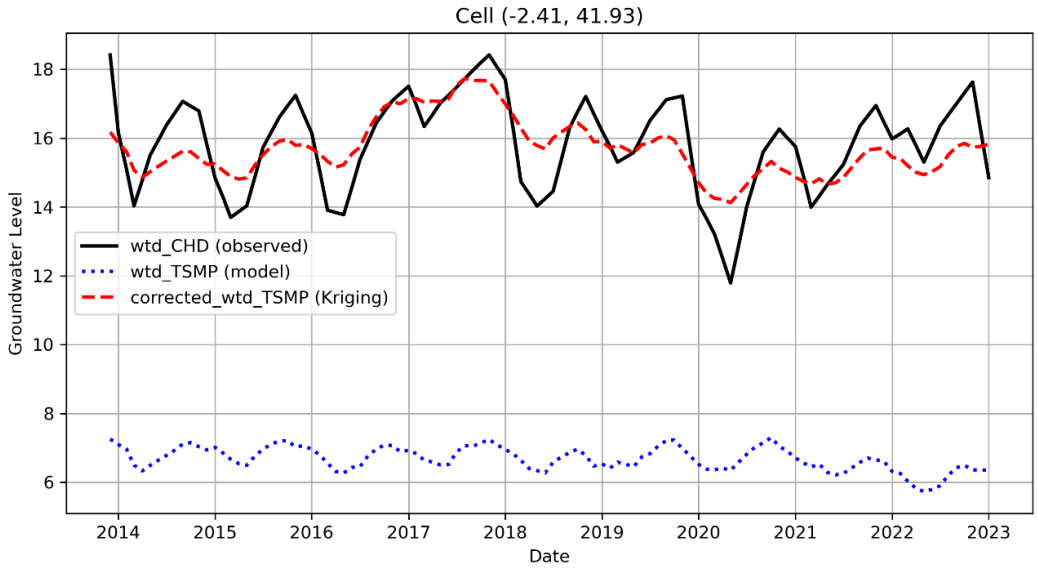


Figure 34. Sample plot of the observed (WTD_CHD), simulated (WTD_TSMP) and corrected time series of water table depth (m) through spatiotemporal kriging in one TSMP grid cell.

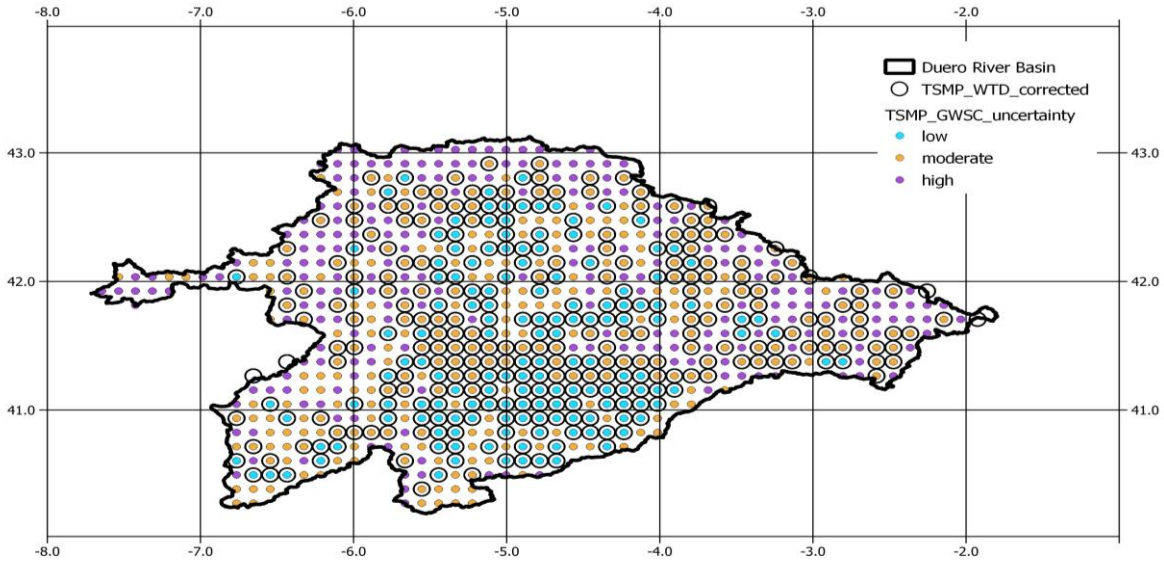


Figure 35. Map of uncertainty in groundwater storage change (GWSC) estimations at the 11 km TSMP pixels of the Duero basin. Open circles are used to identify pixels with available correct water table depth (WTD) data.

Best performing models

Among the tested DL architectures for spatiotemporal groundwater resources estimation, STT has emerged as the most promising surrogate. We implemented a STT to estimate GWSC across space and time.

The modelling system processes 120 months of geospatial data from 2013 to 2022, integrating dynamic environmental features such as precipitation, evapotranspiration, and temperature with static landscape characteristics including land use patterns, soil properties, and geological features

(Table 11). In a first approach, to ensure the forecasting capabilities of the models for future scenarios, only those dynamic features that are available from different climate change scenario datasets have been included: precipitation, potential evapotranspiration and maximum temperature.

The core of the system is a STT neural network that uses self-attention mechanisms to capture complex spatial and temporal dependencies in the data, taking 48 months of historical observations to forecast the next 12 months of groundwater conditions. The STT model employs multi-head self-attention layers with residual connections and positional encoding to capture both spatial and temporal patterns in the data.

The training pipeline begins with sophisticated data preprocessing that handles irregular spatial sampling within the Duero basin boundary, applies robust normalization techniques including power transform for negative GWSC values, and incorporates uncertainty estimates from CLM and TSMP corrections to weight sample importance. The system employs automated hyperparameter optimization using spatial k-fold cross-validation, systematically exploring configurations for the transformer architecture including hidden dimensions (128-384), attention heads (4-12), and network depth (2-6 layers). Training incorporates advanced techniques such as learning rate scheduling, gradient clipping, early stopping, and a custom loss function that combines mean squared error with mean absolute error weighted by data uncertainty.

Uncertainty derived from TWS corrections is both an input feature (so the STT can adapt its attention) and a weight in the loss (so unreliable samples contribute less to training). This two-pronged approach addresses uncertainty at the representation level and the optimisation level.

In the following, we present the results for GWSC prediction of the TSMP data. Beyond the primary STT model, we implement a complementary XGBoost ensemble for comparison and create an optimally weighted hybrid model that combines predictions from both approaches. A summary of model performance for each model is presented in Table 12.

Table 12. Model performance across each model: optimised SpatioTemporal Transformer (STT), XGBoost; and STT-XGBoost ensemble; for the train, validation (val.) and test data. Metrics include Mean Square Error (MSE), Mean Absolute Error (MAE) and R^2.

Model	**STT**			**XGBoost**			**STT-XGBoost ensemble**		
	Train	Val.	Test	Train	Val.	Test	Train	Val.	Test
MSE	0.0015	0.0021	0.0027	0.0000	0.0028	0.0029	0.0010	0.0021	0.0027
MAE	0.0223	0.0278	0.0333	0.0048	0.0335	0.0363	0.0184	0.0282	0.0332
R^2	0.0333	0.4479	0.3847	0.9878	0.2887	0.3418	0.7445	0.4589	0.4021

Figure 36 shows the smoothed training and validation loss curves of the final STT, tuned over and optimization process with 3-folds spatial cross-validation and 30 trials. The temporal data split was set as follows: months 0-47 (train), 48-95 (validation), 96-107 (test). The model achieved moderate but robust performance (Table 12).

Spatial goodness-of-fit maps help visualizing the spatial performance of the STT model across the Duero basin (Figure 37). TWS data from TSMP was only available until 2018, and only simulated WTD were available from 2018 to 2022; that explains the decrease in number of samples in the validation and test sets compared to the train set. The training set covers the entire area, ensuring sufficient information for the STT to learn spatiotemporal patterns in the data. The moderate spatial performance in the train set reflects the different regularization mechanisms used to prevent overfitting. The validation and test maps evidence the effect of the uncertainty layer: better results in areas with less uncertainty in GWSC corrections (Figure 35 and Figure 37).

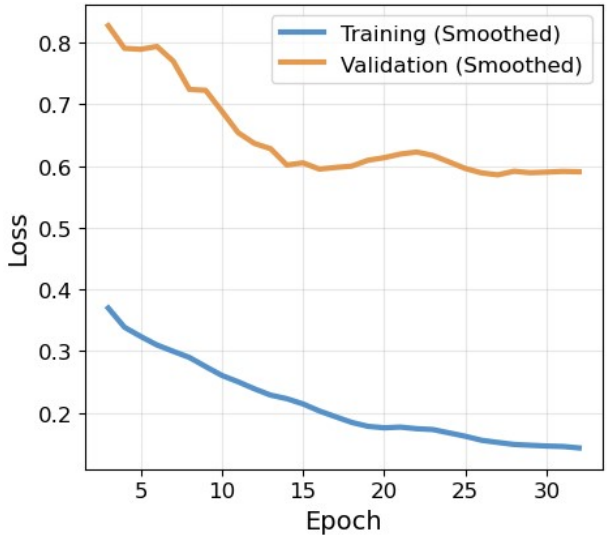


Figure 36. Smoothed training and validation curves of the training process of the optimised spatiotemporal Transformer for groundwater storage change prediction in the Duero River Basin.

Feature importance was analysed using permutation testing. These are the top 10 important static features:

1. Hydrogeological domains from the surface horizon (i.e., related to shallow unconfined groundwater bodies)
2. Population-based water supply use.
3. Maximum volume for hydroelectrical power plants.
4. Groundwater bodies from the lower horizon (i.e., related to deeper groundwater systems from which most water is pumped)
5. Lakes and reservoirs.
6. CORINE land use 216: Heterogeneous Agricultural Areas.
7. SoilGrids: silt content.
8. SoilGrids: sand content.
9. Vegetation types.
10. Maximum volume for livestock farming.

The features are diverse, and they are all highly connected to groundwater dynamics in the Duero basin. The first two features show one order of magnitude higher permutation importance than the rest.

The three dynamic features (precipitation, potential evapotranspiration and maximum temperature) have larger average importance than the static ones, being precipitation and potential evapotranspiration the ones with the highest scores.

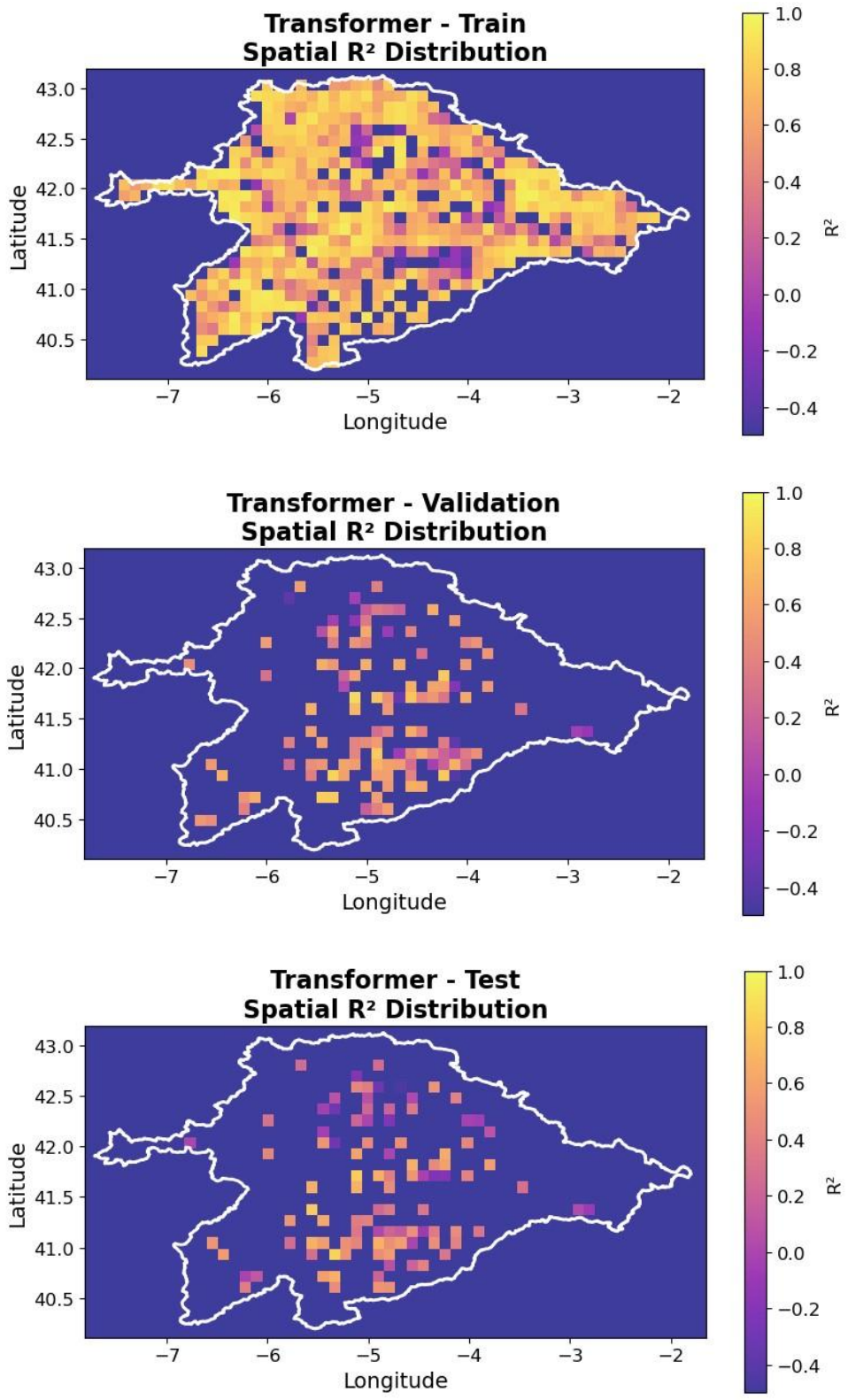


Figure 37. Maps of the spatial distribution of R^2 values for the a) train, b) validation and c) test sets of optimised Spatiotemporal Transformer (STT) across the TSMP pixels in the Duero River Basin. Note the reduction in the number of samples from train (620) to validation (153) and test (153) due to a decrease in available simulated TSMP data from 2018.

The XGBoost model was trained as a multi-output regressor in which dynamic variables were summarised using statistics such as mean, standard deviation, minimum, maximum, and simple linear trend, and concatenated to the static features. This model also showed suitable results, but was prone to overfitting (Table 12).

The STT-XGBoost ensemble based on linear weight search yielded the best metrics (Table 12). The performance improvement is evidenced in the validation and test spatial R^2 maps (Figure 38) compared to the STT alone (Figure 37).

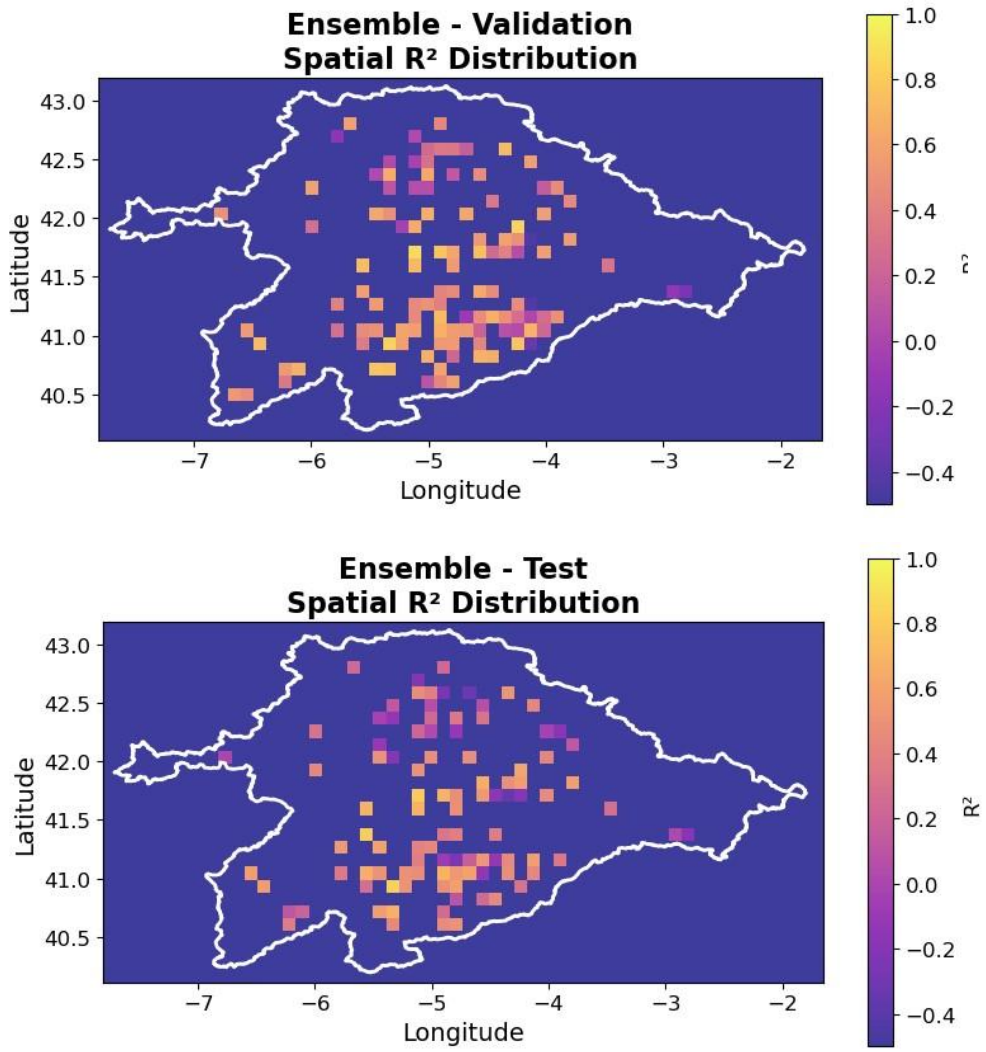


Figure 38. Maps of the spatial distribution of R^2 values for the a) validation and b) test sets of spatiotemporal Transformer (STT) - XGBoost ensemble across the TSMP pixels in the Duero River Basin.

Finally, the time series predictions of the different methods on the top pixels by R^2 were compared in Figure 39. This allows visual inspection of how well each model tracks the actual data at representative locations. The GWSC patterns are captured quite well, especially by the STT alone and the STT-XGBoost ensemble, although peaks and valleys still need to be improved.

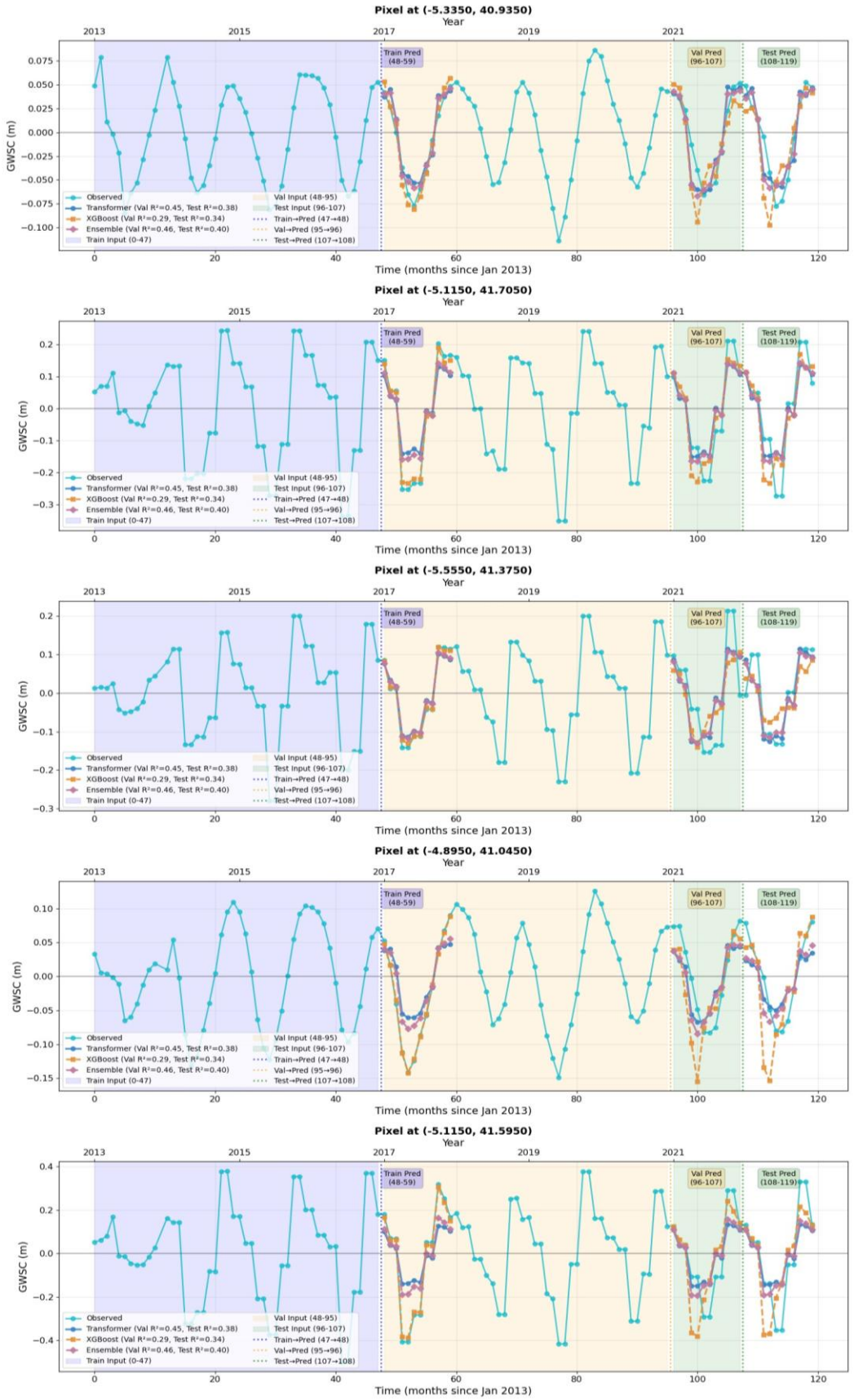


Figure 39. Comparison of predicted time series at pixels with the highest accuracy (R^2). For each selected location, the plots show the observed target series (as points/line), and the predictions from each model: STT (Transformer, dark blue), XGBoost (orange), Ensemble (purple) as dashed lines.

6.4 Discussion

This chapter presents a practical and well-structured approach for producing high-resolution, spatiotemporal estimates of GWSC for the Duero basin. It effectively addresses the limitations of traditional process-based models, which are often computationally intensive and less reliable in data-scarce regions. The core of the methodology is the development of a sophisticated data-driven surrogate model that combines outputs from process models like TSMP and CLM with in-situ groundwater observations. Through geostatistical conditioning techniques such as spatiotemporal kriging and linear correction, a more reliable target variable was created to train advanced machine learning models, primarily a STT complemented by an XGBoost baseline and a final linear STT-XGBoost ensemble.

The framework's design demonstrates several notable strengths. A key contribution is the hybrid data conditioning, which integrates TSMP/CLM outputs with local piezometer data to produce more realistic GWSC targets and a geographically explicit uncertainty flag, a practical method for fusing models and observations for ML training. Furthermore, the study leverages a modern STT architecture within a technically sound workflow that includes robust preprocessing, automated hyperparameter tuning via spatial k-fold cross-validation, and advanced training techniques like early stopping and gradient clipping. A particularly innovative aspect is the uncertainty-aware training, where uncertainty is handled both as an input feature and as a sample weight in the loss function, focusing the learning on more reliable data. This is complemented by a blending of the STT and XGBoost models, which improved validation and test R^2 values, demonstrating the complementary strengths of these different approaches.

Despite these strengths, the approach has several limitations. The ML surrogate is trained on corrected model outputs rather than purely independent observational targets, meaning any biases from the correction or kriging steps could propagate into the final model and its uncertainty flags. This risk is compounded by spatial and temporal sampling gaps; for instance, available TSMP data ends earlier than the target period, reducing the number of samples for validation and testing after 2018 and potentially leading to optimistic performance maps where training data was denser. Additionally, the uncertainty characterization, while pragmatic, is partial; the discrete "low/moderate/high" flags do not fully quantify predictive uncertainty or account for model structural uncertainty.

To build on this promising work, several steps are recommended for future development. The pipeline should be extended to quantify predictive uncertainty probabilistically by producing predictive intervals through methods like quantile regression or deep ensembles, moving beyond the current discrete flags. It is also crucial to validate the correction assumptions and variogram sensitivity through targeted tests to ensure the kriging process introduces minimal bias. Further ablation studies and expanded interpretability analyses would clarify which components drive model skill, building on the initial permutation importance results.

To raise confidence for operational use, transferability tests on different basins should be conducted to evaluate how well the learned patterns generalize. Where possible, enriching the training targets with additional in-situ wells or alternative observational products would reduce the reliance on corrected-model-only targets. For operational monitoring, a continuous update loop should be designed to ingest new observations and retrain the model periodically. Finally, because these outputs will inform water management, it is essential to accompany maps with explicit guidance where high uncertainty implies low decision-making confidence. Overall, this work presents a surrogate modelling

pathway that meaningfully fuses process models and observations, with clear steps to improve bias, uncertainty quantification, and operational confidence.

6.5 Conclusion and next steps

This section presents the framework and results for the development of DL models for GWSC estimation. Due to the lack of reliable GWS observations, WTD time series from the TSMP simulation model were corrected at the pixel level using local WTD observations by means of spatio-temporal kriging and then translated into GWSC using linear approximations of the storage coefficient, upon goodness-of-fit thresholds. So far, the developments show that the corrections presented in this section improve the representativeness of the ML model in the region by incorporating observational data, thereby reducing bias in GWS estimation.

The advanced modelling framework represents a significant step forward in groundwater forecasting, combining state-of-the-art DL techniques with robust geospatial data processing to deliver accurate, interpretable predictions essential for sustainable water resource management. The system's modular design allows for integration with existing hydrological monitoring infrastructure while maintaining flexibility to incorporate additional data sources as they become available.

The outcomes of this initiative are anticipated to yield indicators concerning GWSC and recharge, invaluable for effective groundwater resources management. Nevertheless, as soon as the models are also finalized for the CLM dataset, an open discussion with stakeholders regarding the usefulness of the products will be undertaken.

7 Conclusions

In this report, a set of data-driven modelling tools are described which have been developed to help address the needs of the 7 STARS4Water RBHs. These tools cover a range of water resource related topics, including: reservoirs, agricultural water use, groundwater resources and groundwater quality. Several of these tools have shown promising results for applications in the RBHs, and with continued collaboration with the basin stakeholders have the potential to be deployed operationally to aid water management decisions. Generally, each tool has been applied in one or more RBH but also has the capacity to be extended to other regions – this will be explored further in Task 4.4. The outcomes of this work are summarised here.

LSTM models for reservoir inflow and storage simulations

An LSTM-based multi-task model was used to model reservoir inflows and volumes on a daily timestep in the Duero and Seine RBHs. For the four reservoirs in the Seine, simulations of inflow and storage were promising. This is in contrast to the results for the Camporredondo reservoir (Duero) which simulated inflow well but did not successfully simulate storage, likely due to a lack of data on reservoir operations and low volumes of training data overall. Various data augmentation methods were trialled to address this scarcity which resulted in small improvements in model performance.

Ensemble tree model for reservoir storage forecasting

A range of ensemble tree models were explored for simulating and forecasting reservoir storage on a monthly timestep in the East Anglia and Duero RBHs. Results for storage simulations were good at 1 and 3 month lead times for the majority of reservoirs, with a multi-reservoir model outperforming single-reservoir models in most cases. The multi-reservoir model was also shown to have skill in forecast mode for reservoirs in the UK when run with an ensemble of meteorological forecasts. Future steps include a spatial expansion of the model to other regions, and further analysis of forecast outputs with an aim to produce operational forecasts for reservoir storage.

Estimation of water table depth anomalies

Monthly WTDA were estimated by downscaling GRACE satellite data in the Seine River Basin using RF and LSTM models. Model outputs were evaluated against TSMP simulations and in-situ groundwater observations. The models were shown to successfully emulate the TSMP simulations, thus downscaling global satellite data in a computationally efficient manner, though limitations were evident in areas influenced by coastal processes and karst systems where global datasets lack sufficient resolution. Model performance was varied when compared to in-situ observations, emphasising the need for hybrid approaches that integrate local hydrogeological data with global datasets to enhance accuracy.

Agricultural water use

A RF model was built to estimate irrigated area at 1 km spatial resolution in the Rhine basin, as a first step to quantify the impacts of climate change on agricultural water use. High resolution true-colour and thermal imagery from the Landsat 7 and 8 satellites was used to estimate irrigated area, with the results evaluated against Eurostat statistics at NUTS level 2. The model results align well with subnational statistics, although it performs better in regions with larger agricultural holdings, and as such is a suitable method for estimating irrigated area which can then be used to calculate agricultural water demand.

Predictive mapping of groundwater quality

Various ensemble tree algorithms, implemented through the MLMapper tool, were used to produce predictive maps for nitrate contamination in groundwater for the Duero and East Anglia RBHs. The models performed well when compared to observed data and to prior knowledge. Explainable ML techniques were used to further investigate the models and demonstrated that the models are consistent with conceptual models of groundwater contamination. This method can help water users to better understand groundwater contamination in different boreholes, as well as in deciding where to site new monitoring boreholes. It can be expanded to new areas and other contaminants, provided there is sufficient data available to train and test the algorithms.

Quantitative groundwater resources estimation

A modelling framework was developed to estimate GWSC at fine spatiotemporal scales, and applied to the Duero basin. This framework includes hybrid data conditioning, used to integrate model outputs from TSMP and CLM with local GWL data, to create more realistic GWS targets and a geographically explicit uncertainty flag. The improved data is then used to train DL models, including the uncertainty information, which allows the model to focus the learning on more reliable data. These models produce accurate, interpretable predictions essential for sustainable water resource management, as required by basin stakeholders. This framework maintains flexibility, thus allowing transfer to other basins, and the ability to incorporate additional data sources as they become available.

Overview

Overall, multiple new data-driven tools have been developed and tested in the STARS4Water RBHs to address a range of water related topics. Each of these tools seeks to provide new model functionality, leveraging the capabilities of data-driven models to support water resource management and improve our understanding of the issues. The results of this work underscore ML's growing role as a complementary tool in hydrological sciences, capable of bridging the gap between large-scale datasets and local water management needs. These models and their applications can be used to support the STARS4Water stakeholders and the wider water community, as they strive for adaptive, resilient and sustainable management of freshwater resources.

Bibliography

Abbaszadeh, P., Moradkhani, H., & Zhan, X. (2019). Downscaling SMAP radiometer soil moisture over the CONUS using an ensemble learning method. *Water Resources Research*, 324-344.

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

Agarwal, V., Akyilmaz, O., Shum, C. K., Feng, W., Yang, T.-Y., Forootan, E., Syed, T. H., Haritashya, U. K., & Uz, M. (2023). Machine learning based downscaling of GRACE-estimated groundwater in Central Valley, California. *Science of the Total Environment, 865*, 161138.

Ahmad, S. K., & Hossain, F. (2019). A web-based decision support system for smart dam operations using weather forecasts. *Journal of Hydroinformatics*, 687-707.

Akter, A., & Ahmed, S. (2011). Modeling of groundwater level changes in an urban area. *Sustain Water Resour*, 1-20.

Aller, L., Lehr, J. H., Petty, R., & Bennett, T. (1987). DRASTIC—A Standardized System to Evaluate Groundwater Pollution Potential Using Hydrogeologic Setting. *Journal of the Geological Society of India*, 23-37.

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., & Mai, J. (2023). Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences, 27*(1), 139-157.

Atkinson, P. M. (2013). Downscaling in remote sensing. *International Journal of Applied Earth Observation and Geoinformation, 22*, 106–114.

Avila, L., de Lavenne, A., Ramos, M. H., & Kollet, S. (2025). Estimation of Monthly Water Table Depth Anomalies Based on the Integration of GRACE and ERA5-Land with Large-Scale Simulations Using Random Forest and LSTM Networks. *Water Resources Management*, 1-20.

Bai, J., Cui, Q., Zhang, W., & Meng, L. (2019). An approach for downscaling SMAP soil moisture by combining Sentinel-1 SAR and MODIS data. *Remote Sensing*, 2736.

Baup, F., Frappart, F., & Maubant, J. (2014). Combining High-Resolution Satellite Images and Altimetry to Estimate the Volume of Small Lakes. *Hydrol. Earth Syst. Sci*, 2007-2020.

Ben-Salem, N., Reinecke, R., Copty, N. K., Gómez-Hernández, J. J., Varouchakis, E. A., Karatzas, G. P., & ... & Jomaa, S. (2023). Mapping steady-state groundwater levels in the Mediterranean region: The Iberian Peninsula as a benchmark. . *Journal of Hydrology*, 130207.

Busschaert, L., de Roos, S., Thiery, W., Raes, D., & De Lannoy, G. J. (2022). Net irrigation requirement under different climate scenarios using AquaCrop over Europe. *Hydrology and Earth System Sciences, 26*(14), 3731-3752.

Calla, O. P., Bohra, D., Vyas, R., Purohit, B. S., Prasher, R., Loomba, A., & Kumar, N. (2008). Measurement of soil moisture using microwave radiometer. *2008 International Conference on Recent Advances in Microwave Theory and Applications* (pp. 621-624). IEEE.

Castillo-Botón, C., Casillas-Pérez, D., Casanova-Mateo, C., Moreno-Saavedra, L. M., Morales-Díaz, B., Sanz-Justo, J., Gutiérrez, P. A., & Salcedo-Sanz, S. (2020). Analysis and Prediction of Dammed Water Level in a Hydropower Reservoir Using Machine Learning and Persistence-Based Techniques. *Water*, 1528.

CEDEX. (2024). *C.H. DUERO.*. Retrieved 06 22, 2023, from El Centro de Estudios Hidrográficos. https://ceh.cedex.es/anuarioaforos/DUERO_csv.asp

Chen, C., Chen, Q., Qin, B., Zhao, S., & Duan, Z. (2020). Comparison of different methods for spatial downscaling of GPM IMERG V06B satellite precipitation product over a typical arid to semi-arid area. *Frontiers in Earth Science, 8*, 536337.

Chen, J. L., Wilson, C. R., Tapley, B. D., Yang, Z. L., & Niu, G.-Y. (2009). 2005 drought event in the Amazon River basin as measured by GRACE and estimated by climate models. *Journal of Geophysical Research: Solid Earth, 114*.

Chen, Z., Zheng, W., Yin, W., Li, X., Zhang, G., & Zhang, J. (2021). Improving the spatial resolution of GRACE-derived terrestrial water storage changes in small areas using the Machine Learning Spatial Downscaling Method. *Remote Sensing, 13*, 4760.

Choong, S. M., & El-Shafie, A. (2015). State-of-the-Art for Modelling Reservoir Inflows and Management Optimization. *Water Resources Management*, 1267-1282.

Coerver, H. M., Rutten, M. M., & van de Giesen, N. C. (2018). Deduction of reservoir operating rules for application in global hydrological models. *Hydrology and Earth System Sciences*, 831-851.

Condon, L., Kollet, S., Bierkens, M., Fogg, G., Maxwell, R., Hill, M., Fransen, H.-J., Verhoef, A., Van Loon, A., Sulis, M., & al., e. (2021). Global groundwater modeling and monitoring: opportunities and challenges. *Water Resour Res* .

Delaigue, O., Guimarães, G. M., Brigode, P., Génot, B., Perrin, C., Soubeyroux, J.-M., Janet, B., Addor, N., & Andréassian, V. (2024). CAMELS-FR dataset: A large-sample hydroclimatic dataset for France to explore hydrological diversity and support model benchmarking. *Earth System Science Data* .

Döll, P., & Siebert, S. (2002). Global modeling of irrigation water requirements. *Water Resources Research, 38*(4), 8-1-8-10.

Donchyts, G., Winsemius, H., Baart, F., Dahm, R., Schellekens, J., Gorelick, N., Iceland, C., & Schmeier, S. (2022). High-resolution surface water dynamics in Earth's small and medium-sized reservoirs. *Sci Rep*, 13776.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., Mistelbauer, T., & Lecomte, P. (2017). ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment*, 185-215.

Droppers, B., Leijnse, M., Bierkens, M. F., & Wanders, N. (2023). Introducing DL-GLOBWB: a deep-learning surrogate of a process-based global hydrological model. Vienna, Austria: EGU, Copernicus Meetings.

Du, Y., Zhang, Y., Ling, F., Wang, Q., Li, W., & Li, X. (2016). Water Bodies' Mapping from Sentinel-2 Imagery with Modified Normalized Difference Water Index at 10-m Spatial Resolution Produced by Sharpening the Swir Band. *Remote Sens.*, 354.

Durant, M., & Counsell, C. (2018). *Inventory of reservoirs amounting to 90% of total UK storage*. NERC Environmental Information Data Centre. https://doi.org/10.5285/f5a7d56c-cea0-4f00-b159-c3788a3b2b38

Ehsani, N., Fekete, B. M., Vörösmarty, C. J., & Tessler, Z. D. (2016). A neural network based general reservoir operation scheme. *Stochastic environmental research and risk assessment*, 1151-1166.

EPTB SGL. (2023). Daily time series of reservoir data for 4 reservoirs in the Seine. [Unpublished dataset].

Famiglietti, J. S., Lo, M., Ho, S. L., Bethune, J., Anderson, K. J., Syed, T. H., Swenson, S. C., de Linage, C. R., & Rodell, M. (2011). Satellites measure recent rates of groundwater depletion in California's Central Valley. *Geophysical Research Letters, 38*.

Fasbender, D., & Ouarda, T. B. (2010). Spatial Bayesian model for statistical downscaling of AOGCM to minimum and maximum daily temperatures. *Journal of climate, 23*, 5222–5242.

Ferro, C. A. (2014). Fair scores for ensemble forecasts. *Q. J. R. Meteorol. Soc.*, 1917-1923.

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.

Gaur, S., Chahar, B. R., & Graillot, D. (2011). Combined use of groundwater modeling and potential zone analysis. *Int J Appl Earth Obs Geoinf*, 127–139.

Gleick, P. H. (2003). Global freshwater resources: soft-path solutions for the 21st century. *Science*, 1524-1528.

Gleick, P. H., Cooley, H., Famiglietti, J. S., Lettenmaier, D. P., Oki, T., Vörösmarty, C. J., & Wood, E. F. (2013). Improving understanding of the global hydrologic cycle. . In *Climate science for serving society.* (pp. 151-184).

Gómez-Escalonilla, V., & Martínez-Santos, P. (2024). A machine learning approach to map the vulnerability of groundwater resources to agricultural contamination. *Hydrology*, 153.

Gómez-Escalonilla, V., Diancoumba, O., Traoré, D. Y., Montero, E., Martín-Loeches, M., & Martínez-Santos , P. (2022). Multiclass spatial predictions of borehole yield in southern Mali by means of machine learning classifiers. *Journal of Hydrology: Regional Studies, 44*, 101245.

Gómez-Escalonilla, V., Diancoumba, O., Traoré, D. Y., Montero, E., Martín-Loeches, M., & Martínez-Santos, P. (2022a). Multiclass spatial predictions of borehole yield in southern Mali by means of machine learning classifiers. *Journal of Hydrology: Regional Studies, 44*, 101245.

Gómez-Escalonilla, V., Martínez-Santos, P., De la Hera-Portillo, A., Díaz-Alcaide, S., Montero, E., & Martín-Loeches, M. (2024a). A machine learning application for the development of groundwater vulnerability studies. *15th International Conference of Hydroinformatics ,Beijing, China*, May 27-30 2024.

Gómez-Escalonilla, V., Martínez-Santos, P., Díaz-Alcaide, S., Montero, E., & Martín-Loeches, M. (. (2024b). GIS-based machine learning applications as decision support systems to enhance groundwater monitoring networks. *15th International Conference of Hydroinformatics. Beijing; China.*, May 27-30 2024.

Gómez-Escalonilla, V., Martínez-Santos, P., Pacios, D., Ruíz-Álvarez, L., Díaz-Alcaide, S., Montero, E., Martín- Loeches, M., De la Hera-Portillo, A., & Aguilera, H. (2024c). Nitrate spatial predictions by means of machine learning to improve groundwater monitoring networks. *EGU General Assembly*, 14–19 April 2024.

Gómez-Escalonilla, V., Vogt, M. L., Destro, E., Isseini, M., Origgi, G., Djoret, D., Martínez-Santos, P., & Holecz, F. (2022b). Delineation of groundwater potential zones by means of ensemble tree supervised classification methods in the Eastern Lake Chad basin. *Geocarto International, 37*(25), 8924-8951.

Gourgouletis, N., Bariamis, G., Anagnostou, M. N., & Baltas, E. (2022). Estimating Reservoir Storage Variations by Combining Sentinel-2 and 3 Measurements in the Yliki Reservoir, Greece. *Remote Sens.*, 1860.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal oh Hydrology*, 80-91.

Gutenson, J. L., Tavakoly, A. A., Wahl, M. D., & Follum, M. L. (2020). Comparison of generalized non-data-driven lake and reservoir routing models for global-scale hydrologic forecasting of reservoir outflow at diurnal time steps. *Hydrology and Earth System Sciences*, 2711-2729.

Hartick, C. a.-P. (2021). An interannual probabilistic assessment of subsurface water storage over Europe using a fully coupled terrestrial model. *Water resources research*, e2020WR027828.

Hofmann, C., Schmid, S. L., Lehner, B., Klotz, D., & Hochreiter, S. (2024). Energy-based Hopfield Boosting for Out-of-Distribution Detection. *The Thirty-eighth Annual Conference on Neural Information Processing Systems, Vancouver*.

Hollis, D., McCarthy, M., Kendon, M., Legg, T., & Simpson, I. (2019). HadUK-Grid - A new UK dataset of gridded climate observations. *Geosciences*.

Hong, J., Lee, S., Bae, J. H., Lee, J., Park, W. J., Lee, D., Kim, J., & Lim, K. J. (2020). Development and evaluation of the combined machine learning models for the prediction of dam inflow. *Water*, 2927.

Houborg, R., & McCabe, M. F. (2018). Daily Retrieval of NDVI and LAI at 3 m Resolution via the Fusion of CubeSat, Landsat, and MODIS data. *Remote Sensing*, 890.

Houborg, R., & McCabe, M. F. (2018). A Cubesat Enabled Spatio-Temporal Enhancement Method (CESTEM) utilizing Planet, Landsat and MODIS data. *Remote Sensing of Environment*, 211-226.

Hughes M., B. H. (2004). The development of a GIS-based inventory of standing waters in Great Britain together with a risk-based prioritisation protocol. . *Water, Air, and Soil Pollution: Focus*, 73-84.

Ibañez, S. C., Dajac, C. V., Liponhay, M. P., Legara, E. F., Esteban, J. M., & Monterola, C. P. (2021). Forecasting reservoir water levels using deep neural networks: A case study of Angat dam in the Philippines. . *Water*, 34.

Iwana, B. K., & Uchida, S. (2021). An empirical survey of data augmentation for time series classification with neural networks. *PLoS ONE*. https://doi.org/ https://doi.org/10.1371/journal.pone.0254841

Jia, X., Zhu, Y., & Luo, Y. (2017). Soil moisture decline due to afforestation across the Loess Plateau, China. *Journal of Hydrology*, 113-122.

Kaicun Wang, R. E. (2012). A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. *Review of Geophysics*, 50-2.

Keune, J., Sulis, M., & Kollet, S. J. (2019). Potential added value of incorporating human water use on the simulation of evapotranspiration and precipitation in a continental-scale bedrock-to-atmosphere modeling system: A validation study considering observational uncertainty. *Journal of Advances in Modeling Earth Systems, 11*, 1959–1980.

Kim, J., Read, L., Johnson, L. E., Gochis, D., Cifelli, R., & Han, H. (2020). An experiment on reservoir representation schemes to improve hydrologic prediction: coupling the national water model with the HEC-ResSim. *Hydrological Sciences Journal*, 1652-1666.

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980v9*.

Kollet, S. J., & Maxwell, R. M. (2006). Integrated surface-groundwater flow modeling: a free-surface overland flow. *Adv Water Resour*, 945–958.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.*, 5089-5110.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., & Matias, Y. (2023). Caravan - A global community dataset for large-sample hydrology. *Sci Data*, 61.

Künsch, H. R. (1989). The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics*, 1217 - 1241.

Kurtz, W., He, G., Kollet, S. J., Maxwell, R. M., Vereecken, H., & Hendricks Franssen, H.-J. (2016). TerrSysMP–PDAF (version 1.0): a modular high-performance data assimilation framework for an integrated land surface–subsurface model. *Geoscientific Model Development, 9*, 1341–1360.

Lakshmi, V. (2013). Remote sensing of soil moisture. *International Scholarly Research Notices*.

Lall, U., Josset, L., & Russo, T. (2020). A snapshot of the world's groundwater challenges. *Annu Rev Environ*, 171–194.

Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rodel, R., Sindorf, N., & Wisser, D. (2011). High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Frontiers in Ecology and the Environment*, 494-502.

Lin, Y., Li, X., Zhang, T., Chao, N., Yu, J., Cai, J., & Sneeuw, N. (2020). Water Volume Variations Estimation and Analysis Using Multisource Satellite Data: A Case Study of Lake Victoria. *Remote Sens.*, 3052.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., & Thieme, M. (2019). Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data*, 2052-4463.

Liu, Z., Liu, P.-W., Massoud, E., Farr, T. G., Lundgren, P., & Famiglietti, J. S. (2019). Monitoring groundwater change in California's central valley using sentinel-1 and grace observations. *Geosciences, 9*, 436.

Lu, X., Wei, M., Tang, G., & Zhang, Y. (2018). Evaluation and correction of the TRMM 3B43V7 and GPM 3IMERGM satellite precipitation products by use of ground-based data over Xinjiang, China. *Environmental earth sciences, 77*, 1–18.

Maréchal, J.-C. J. (n.d.).

Maréchal, J.-C., & Rouillard, J. (2020). Groundwater in france: resources, use and management issues. In J.-D. Rinaudo, C. Holley, S. Barnett, & M. Montginoul, *Sustainable groundwater management: a comparative analysis of French and Australian policies and implications to other countries.* (pp. 17-45). Heidelberg: Springer.

Martínez-Santos, P., & Renard, P. (2019). Mapping Groundwater Potential Through an Ensemble of Big Data Methods. *Groundwater, 58*(4), 583-597.

Martínez-Santos, P., Aristizábal, H. F., Díaz-Alcaide, S., & Gomez-Escalonilla, V. (2021). Predictive mapping of aquatic ecosystems by means of support vector machines and random forests: an application to the Valle del Cauca region, Colombia. *Journal of Hydrology, 595*, 126026.

Mignolet, C., Schott, C., & Benoît, M. (2007). Spatial dynamics of farming practices in the Seine Basin: methods for agronomic approaches on a regional scale. *Sci Total Environ*, 13-32.

Miro, M. E., & Famiglietti, J. S. (2018). Downscaling GRACE remote sensing datasets to high-resolution groundwater storage change maps of California's Central Valley. *Remote Sensing, 10*, 143.

Müller, S., Schüler, L., Zech, A., & Heße, F. (2022). GSTools v1.3: a toolbox for geostatistical modelling in Python. *Geosci. Model Dev.*, 3161–3182.

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. Technical Report ECMWF Tech. Memo.* https://doi.org/10.21957/qq4v2o7oy.10.21957: ECMWF.

Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*.

NRFA. (2022, 03 18). *Monthly Hydrological Summaries*. National River Flow Archive . https://nrfa.ceh.ac.uk/monthly-hydrological-summary-uk

Pacios, D., Coleto, I., Verzier, P., Gómez-Escalonilla, V., & Martínez-Santos , P. (2023). Machine learning as a tool to improve groundwater monitoring networks. *50th IAH Congress.* Cape Town, South Africa: Groundwater Division.

Pedregosa, F., Varoquaux , G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*.

Peng, J., Loew, A., Merlin, O., & Verhoest, N. E. (2017). A review of spatial downscaling of satellite remotely sensed soil moisture. *Reviews of Geophysics, 55*, 341–366.

Peñuela, A., Hutton, C., & Pianosi, F. (2020). Assessing the value of seasonal hydrological forecasts for improving water resource management: insights from a pilot application in the UK. *Hydrology and Earth System Sciences*, 6059–6073.

Prewitt, J. M., & Mendelsohn, M. L. (2006). The Analysis of Cell Images. *Ann. N. Y. Acad. Sci.*, 1035-1053.

Pulla, S. T., Yasarer, H., & Yarbrough, L. D. (2023). GRACE Downscaler: A Framework to Develop and Evaluate Downscaling Models for GRACE. *Remote Sensing, 15*(9), 2247.

Purnamasari, D., Teuling, A., & Weerts, A. (2025). Identifying irrigated areas using land surface temperature and hydrological modelling: application to the rhine basin. *Hydrology and Earth System Sciences, 29 (6)*, 1483–1503.

Raghavendra, N. S., & Deka, P. C. (2014). Support vector machine applications in the field of hydrology: A review. *Applied soft computing*, 372-386.

Rahman, M., Rosolem, R., Kollet, S., & Wagener, T. (2019). Towards a computationally efficient free-surface ground-. *Adv Water Resour* , 225-233.

Rodell, M., Velicogna, I., & Famiglietti, J. S. (2009). Satellite-based estimates of groundwater depletion in India. *Nature, 460*, 999–1002.

Rousset, F., Habets, F., Gomez, E., Le Moigne, P., Morel, S., Noilhan, J., & Ledoux, E. (2004). Hydrometeorological modeling of the Seine basin using the SAFRAN-ISBA-MODCOU system. *Journal of Geophysical Research Atmosphere*.

Saadi, S., Todorovic, M., Tanasijevic, L., Pereira, L. S., Pizzigalli, C., & Lionello, P. (2015). Climate change and Mediterranean agriculture: Impacts on winter wheat and tomato crop evapotranspiration, irrigation requirements and yield. *Agricultural Water Management, 147*, 103-115.

Sabzehee, F., Amiri-Simkooei, A. R., Iran-Pour, S., Vishwakarma, B. D., & Kerachian, R. (2023). Enhancing spatial resolution of GRACE-derived groundwater storage anomalies in Urmia catchment using machine learning downscaling methods. *Journal of Environmental Management, 330*, 117180.

Sapitang, M. M., Ridwan, W., Faizal Kushiar, K., Najah Ahmed, A., & El-Shafie, A. (2020). Machine learning application in reservoir water level forecasting for sustainable hydropower generation strategy. *Sustainability*, 6121.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference.*

Seo, J. Y., & Lee, S. I. (2021). Predicting changes in spatiotemporal groundwater storage through the integration of multi-satellite data and deep learning models. *IEEE Access*, 157571-157583.

Seyoum, W. M., Kwon, D., & Milewski, A. M. (2019). Downscaling GRACE TWSA data into high-resolution groundwater level anomaly using machine learning-based models in a glacial aquifer system. *Remote Sensing, 11*, 824.

Shin, S., Pokhrel, Y., & Miguez-Macho, G. (2018). High resolution modeling of reservoir release and storage dynamics at the continental scale. *Water resources research*.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 2196-1115. https://doi.org/10.1186/s40537-019-0197-0

Shrestha, P., Sulis, M., Masbou, M., Kollet, S., & Simmer, C. (2014). A scale-consistent terrestrial systems modeling platform based on COSMO, CLM, and ParFlow. *Monthly weather review, 142*, 3466–3483.

Srivastava, P. K., Han, D., Ramirez, M. R., & Islam, T. (2013). Machine learning techniques for downscaling SMOS satellite soil moisture using MODIS land surface temperature for hydrological application. *Water resources management*, 3127-3144.

Stringer, N., Knight, J., & Thornton, H. (2020). Improving Meteorological Seasonal Forecasts for Hydrological Modeling in European Winter. *J. Appl. Meteorol. Clim*, 317–332.

Sulis, M., Keune, J., Shrestha, P., Simmer, C., & Kollet, S. J. (2018). Quantifying the impact of subsurface-land surface physical processes on the predictive skill of subseasonal mesoscale atmospheric simulations. *Journal of Geophysical Research: Atmospheres, 123*, 9131–9151.

Tan, S., Wu, B., & Yan, N. (2019). A method for downscaling daily evapotranspiration based on 30-m surface resistance. *Journal of hydrology, 577*, 123882.

Tavakoly, A., Habets, F., Saleh, F., Yang, Z.-L., Bourgeois, C., & Maidment, D. (2019). An integrated framework to model nitrate contaminants with interactions of agriculture, groundwater, and surface water at regional scales: the STICS-EauDyssée coupled models applied over the Seine River Basin. *J Hydrol*, 943-958.

Tiwari, A. D., & Mishra, V. (2019). Prediction of reservoir storage anomalies in India. *Journal of Geophysical Research: Atmospheres*.

Turner, S. W., Bennett, J. C., Robertson, D. E., & Galelli, S. (2017). Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. . *Hydrology and Earth System Sciences.*, 4841–4859.

Turner, S. W., Doering, K., & Voisin, N. (2020). Data-driven reservoir simulation in a large-scale hydrological and water resource model. *Water resources research*.

UKCEH. (2025). *UK Hydrological Outlook*. https://hydoutuk.net/

Valipour, M., Banihabib, M. E., & Behbahani, S. M. (2013). Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *Journal of hydrology*, 433–441.

Wackernagel, H. (2003). *Multivariate geostatistics.* Berlin, Heidelberg: Springer.

Wan, Z., Hook, S., & Hulley, G. (2021). MODIS Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061. *NASA EOSDIS Land Processes Distributed Active Archive Center*.

Wan, Z., Hook, S., & Hulley, G. (2021, 08 11). *MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061 [Data set]*. https://doi.org/10.5067/MODIS/MYD11A1.061. https://doi.org/10.5067/MODIS/MYD11A1.061

Wan, Z., Hook, S., & Hulley, G. (2021). MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061. *NASA EOSDIS Land Processes Distributed Active Archive Center [data set]*. https://doi.org/10.5067/MODIS/MYD11A1.061.

Wan, Z., Hook, S., & Hulley, G. (2025, 08 11). MODIS Aqua Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061. *NASA Land Processes Distributed Active Archive Center*. MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061. https://doi.org/10.5067/MODIS/MYD11A1.061

Wang, F., Tian, D., Lowe, L., Kalin, L., & Lehrter, J. C. (2020). Deep Learning for Daily Precipitation and Temperature Downscaling. *AGU Fall Meeting Abstracts*, *2020*, H178–10.

Wang, F., Wang, L., Zhou, H., Saavedra Valeriano, O. C., Koike, T., & Li, W. (2012). Ensemble hydrological prediction-based real-time optimization of a multiobjective reservoir during flood season in a semiarid basin with global numerical weather predictions. *Water resources research.*, 1-21.

Wang, Q., & Wang, S. (2020). Machine Learning-Based Water Level Prediction in Lake Erie. *Water*, 2654.

Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C., & Landerer, F. W. (2015). Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. *Journal of Geophysical Research: Solid Earth, 120*, 2648–2671.

Wen, F., Zhao, W., Wang, Q., & Sánchez, N. (2019). A value-consistent method for downscaling SMAP passive soil moisture with MODIS products using self-adaptive window. *IEEE Transactions on Geoscience and Remote Sensing, 58*, 913–924.

Wiese, D. N., Landerer, F. W., & Watkins, M. M. (2016). Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution. *Water Resources Research, 52*, 7490–7502.

Wilby, R. L., & Wigley, T. M. (1997). Downscaling general circulation model output: a review of methods and limitations. *Progress in physical geography, 21*, 530–548.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences.* Oxford: Academic Press.

Wisser, D., Frolking, S., Douglas, E. M., Fekete, B. M., Vörösmarty, C. J., & Schumann , A. H. (2008). Global irrigation water demand: Variability and uncertainties arising from agricultural and climate data sets. *Geophysical Research Letters, 35*, L24408.

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. *Sci Rep*, 12:8085.

Yan, X., Chen, H., Tian, B., Sheng, S., Wang, J., & Kim, J.-S. (2021). A downscaling–merging scheme for improving daily spatial precipitation estimates based on random forest and cokriging. *Remote Sensing, 13*, 2040.

Yang, T., Asanjan, A. A., Welles, E., Gao, X., Sorooshian, S., & Liu, X. (2017). Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. *Water resources research*, 2786-2812.

Yang, T., Gao, X., Sorooshian, S., & Li, X. (2016). Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. *Water resources research*, 1626-1651.

Yao, H., Wang, Y., Zhang, L., Zou, J., & Finn, C. (2022). C-Mixup: Improving Generalization in Regression. *Proceeding of the Thirty-Sixth Conference on Neural Information Processing Systems*.

Yassin, F., Razavi, S., Elshamy, M., Davison, B., Sapriza-Azuri, G., & Wheater, H. (2019). Representation and improved parameterization of reservoir operation in hydrological and land-surface models. *Hydrology and Earth System Sciences*, 3735-3764.

Zarei, M., Bozorg-Haddad, O., Baghban, S., Delpasand, M., Goharian, E., & Loáiciga, H. A. (2021). Machine-learning algorithms for forecast-informed reservoir operation (FIRO) to reduce flood damages. *Scientific Reports*, 24295.

Zhang, C., Lv, A., Zhu, W., Yao, G., & Qi, S. (2021). Using Multisource Satellite Data to Investigate Lake Area, Water Level, and Water Storage Changes of Terminal Lakes in Ungauged Regions. *Remote Sens.*, 3221.

Zhang, D., Lin, J., Peng, Q., Wang, D., Yang, T., Sorooshian, S., Liu, X., & Zhuang, J. (2018). Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. *Journal of hydrology*, 720-736.

Zhou, H., Luo, Z., Tangdamrongsub, N., Wang, L., He, L., Xu, C., & Li, Q. (2017). Characterizing drought and flood events over the Yangtze River Basin using the HUST-Grace2016 solution and ancillary data. *Remote Sensing, 9*, 1100.

Zhu, S., Hrnjica, B., Ptak, M., Choiński, A., & Sivakumar, B. (2020). Forecasting of water level in multiple temperate lakes using machine learning models. *Journal of hydrology*, 124819.

Zhu, X., Chen, J., Gao, F., Chen, X., & Masek, J. G. (2010). An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sensing of Environment*, 2610-2623.

Appendix A: Monitoring Embalsa Camporredondo using Planet Fusion

Background

Work Package 3 of the STARS4Water Horizon Europe project is developing methods to support water resources management and climate change mitigation. The methods make use of global datasets including satellite imagery. One of the initiatives is to use satellite imagery for reservoir monitoring.

For many reservoirs around the globe, historical and real-time storage measurement data are readily available from the reservoir operator. The water level is easily measured by a staff gauge or float. The storage volume follows from the Stage-Storage Relationship (Depth-Volume-Area). However, there are cases where such data is not available, for example when the data is considered proprietary to the reservoir management organization, or undisclosed for other reasons (e.g. transboundary rivers). In such cases, a monitoring service based on remotely sensed data may be a solution. Good results have been achieved by combining Sentinel-2 and other publicly available optical imagery (Baup et al., 2014; Du et al., 2016; Gourgouletis et al., 2022; Lin et al., 2020; Zhang et al., 2021). However, as these studies are based on public satellite data, they are limited by the resolution and revisit time of public missions, which is 10 m and once every three days in the case of Sentinel-2. For small reservoirs, reservoirs with steep slopes and for regions that are often covered by clouds, this can lead to uncertainty. In such cases, commercial satellite missions that provide higher resolution data and higher revisit times may be a solution. Specifically, the PlanetScope constellation of 100+ CubeSats in low earth orbits represents a novel observational resource, with unprecedented spatial and temporal resolution.

Within STARS4Water, we are conducting a pilot to investigate the added value of Planet commercial satellite data for reservoir surface area monitoring. In a second step, we explore methods to translate reservoir area into storage volume if the Stage-Storage Relationship is not known. Finally, we investigate how the satellite-based reservoir data can be linked to hydrologic catchment modelling to generate derived data that goes beyond the storage monitoring and offers additional benefits to water resources management.

Method

We adopt the method that is followed by earlier studies, employing the Normalized Difference Water Index (NDWI) to detect surface water. Finding the optimal NDWI threshold value between water and non-water pixels is the main challenge. The thresholding method of Prewitt and Mendelsohn (2006) has yielded the best results so far. The NDWI histogram is smoothed until it has two local maxima and the optimal threshold is identified as the minimum value between them. The method was implemented operationally by Donchyts et al (2022). This application is known as Global Water Watch (https://www.globalwaterwatch.earth/).

For the first pilot, we have selected the Camporredondo reservoir in the Duero basin in Spain, which is one of the STARS4Water focus basins. For this reservoir, daily storage and outflow measurement data are available from October 2014 until June 2023 (CEDEX, 2024). The inflow data can be derived from the changes in storage and daily outflow data (Figure 40).

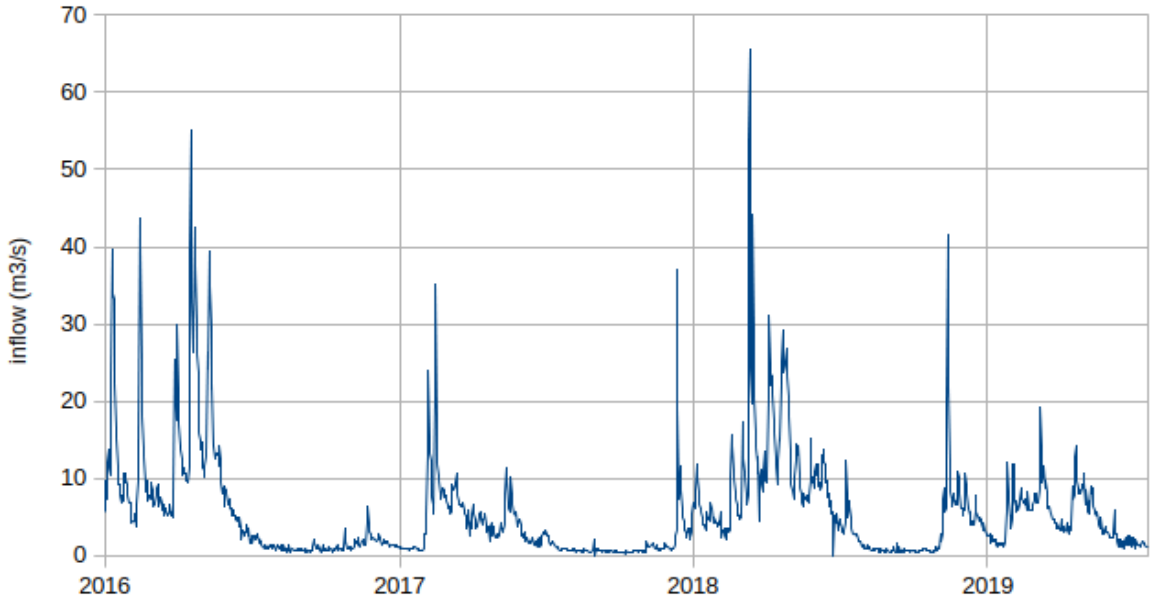


Figure 40. Inflow time series for Camporredondo reservoir, derived from in situ volume and outflow data.

The reservoir surface area measurements were obtained from two sources:

- Global Water Watch, derived from Sentinel-2 and LandSat satellite imagery (https://www.globalwaterwatch.earth/)
- Planet Fusion data, a merged product from PlanetScope and public satellite imagery

Planet has developed a methodology called CubeSat-Enabled Spatio-Temporal Enhancement Method (CESTEM) to enhance, harmonize, inter-calibrate, and fuse cross-sensor data streams (Houborg & McCabe, 2018; Houborg & McCabe, 2018). CESTEM merges publicly accessible multispectral satellites (i.e. Sentinel, Landsat, MODIS) with the higher spatial and temporal resolution data provided by Planet's PlanetScope imagery. The result is a next generation, analysis ready, harmonized Level-3 data product (i.e. maximum amount of available data), which delivers a clean (i.e. free from clouds and shadows), gap-filled (i.e., daily, 3 m), temporally consistent, and radiometrically accurate surface reflectance data product.

Results

Figure 41 shows the detected surface water occurrence of Camporredondo reservoir, derived from Planet Fusion data between January 2022 and July 2023. The deepest parts of the reservoir are near the outlet in the south. The shallower parts in the east and west fall dry more often.

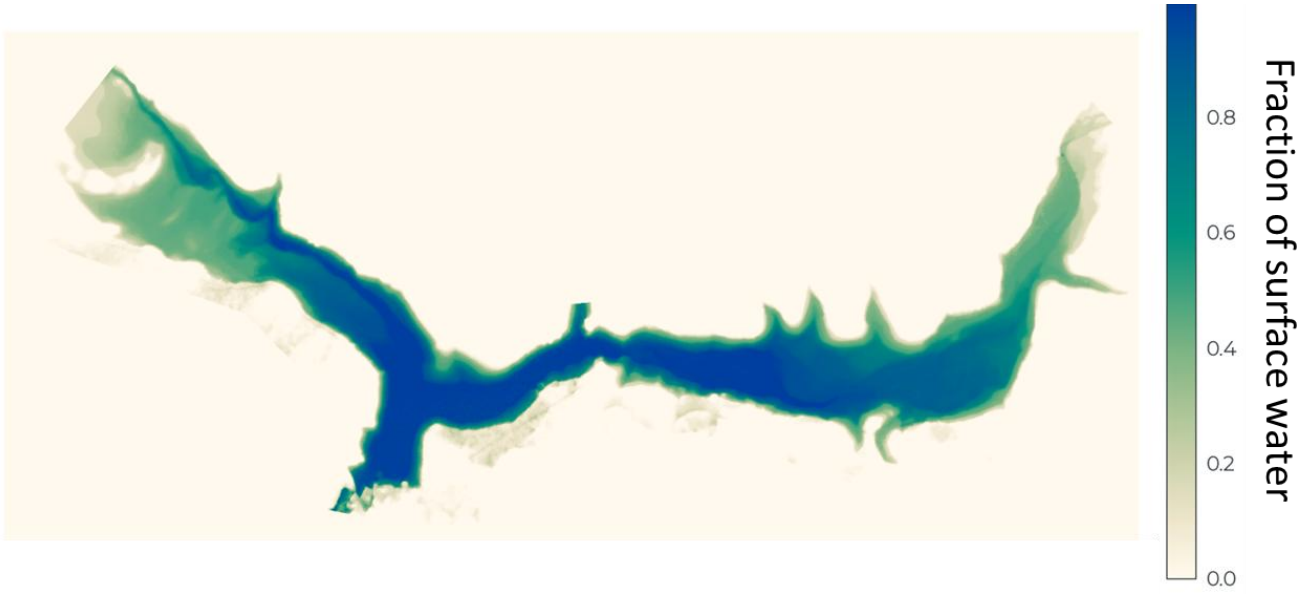


Figure 41. Surface water occurrence in Camporredondo reservoir from Planet Fusion imagery between January 2022 and July 2023.

Figure 42 shows the time series of earth observation-observation (EO)-derived surface area compared to the in-situ storage data. The uncertainty bands in the Planet Fusion graph are based on the per-pixel uncertainty of water detection. The WaterWatch time series is clearly more sparse than that of PlanetFusion.

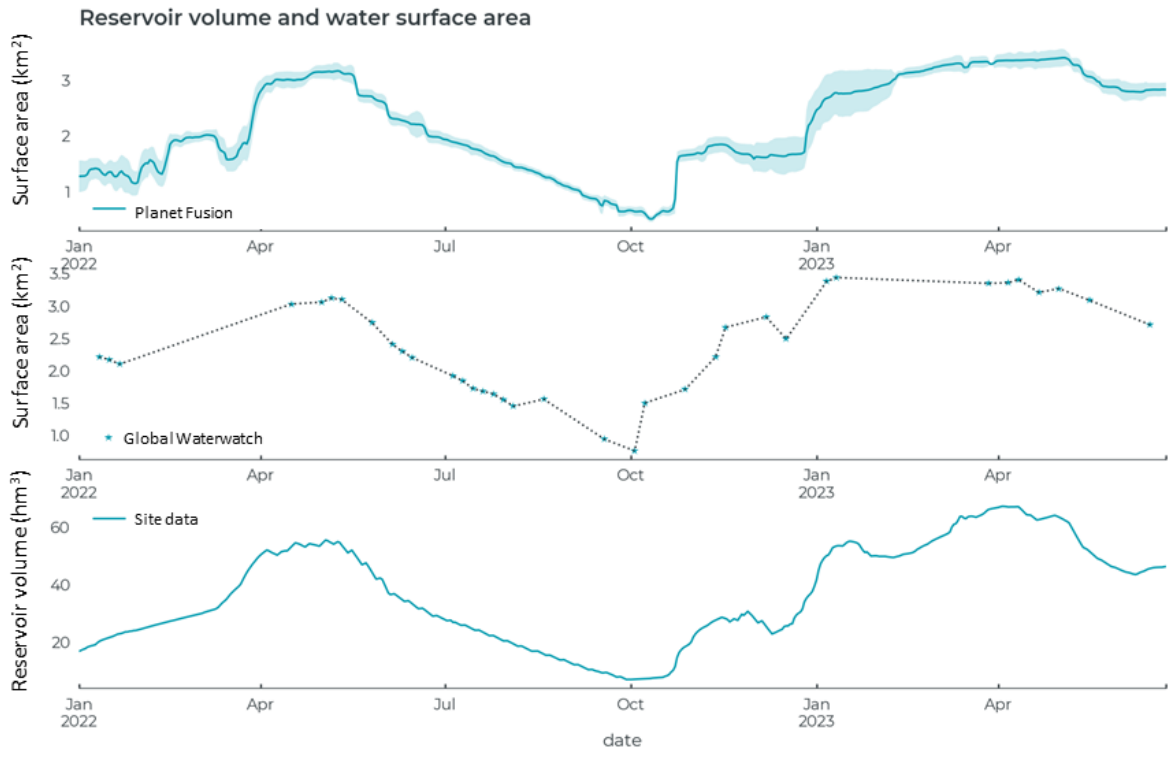


Figure 42. Reservoir surface area, derived from Planet Fusion (top), WaterWatch (middle) and reservoir storage volume from continuous in situ measurements (bottom), between January 2022 and July 2023.

The Stage-Storage Relationship, as derived from the EO-based surface areas and in situ storage data is shown in Figure 43. The Pearson correlation coefficient is 0.92 for the Global Water Watch data. The Planet Fusion correlation is 0.97. From the number of points, it is clear that Planet Fusion yields far more observations than Global Water Watch. This is related to the higher revisit time of the PlanetScope satellites.

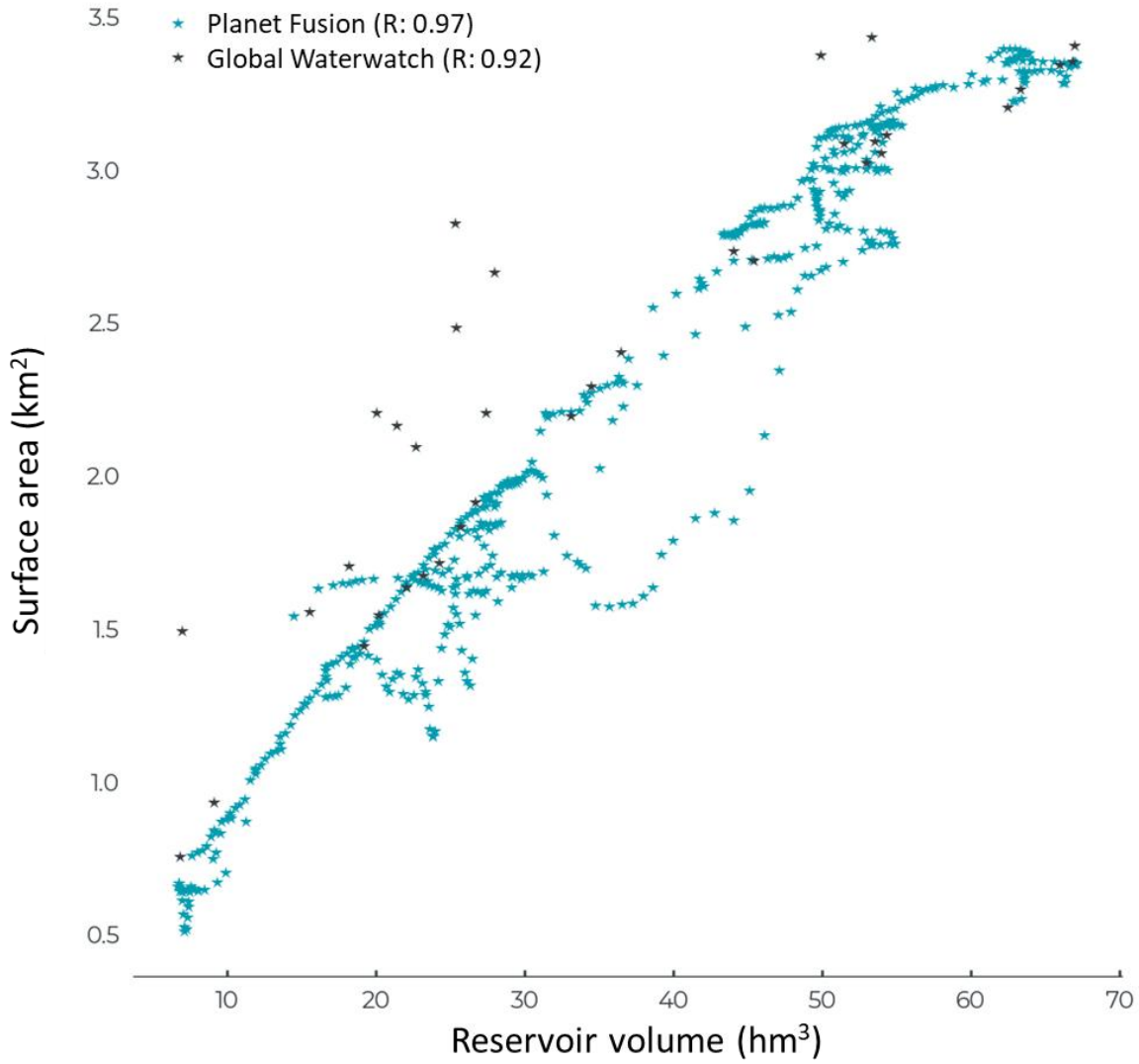


Figure 43. Relationship between estimated reservoir surface area, as derived from Planet Fusion and Global Water Watch, and reservoir storage volume from in situ measurements. The R value, correlation coefficient, is provided for each estimation method.

Interestingly, the deviations from the Stage-Storage Relationship for Global Water Watch tend to be on the positive side, while the Planet Fusion deviations are mostly negative. In other words, Global Water Watch tends to overestimate the surface area, whilst Planet Fusion tends to underestimate it. Further analysis revealed that the overestimation within the WaterWatch method is a result of hazy imagery. As a result, the edges of the reservoir lake are blurred, eventually leading to shoreline pixels being counted as water pixels (Figure 44). The underestimation observed in the Planet Fusion method is due to partly cloudy images that were not filtered out. This results in water pixels not being recognized as water.

Figure 44. Global Water Watch blurred image (right hand side), which leads to an overestimation of reservoir surface area.

Conclusion and next steps

The proof of concept has shown that satellite imagery can be used for Camporredondo reservoir surface area monitoring. The Global Water Watch implementation produces reasonable results, but the higher revisit of PlanetScope leads to an improvement of the Planet Fusion product in terms of a more continuous and consistent signal. Both Global Water Watch and Planet Fusion still contain a few occasional errors and the algorithms can still be optimised further, but the overall performance is considered sufficient for near real time monitoring of reservoir storage (with some uncertainty bounds).

The reservoir surface area can be converted into a reservoir volume by using historical inflow and outflow measurement data. For Camporredondo, these data are available. For other reservoirs, they may not. This will be one of the follow-up investigations. The stage-storage relationship of the reservoir could be derived from a high-resolution digital elevation model or from Sentinel-3 altimetry data. This would enable global reservoir monitoring without any need for in situ calibration data. Another follow-up activity is to employ the storage data in hydrological modelling.

Appendix B: Predicting reservoir storage using ensemble-tree models

The following steps are employed to build the reservoir-specific models:

- **Data selection:** daily storage data for each reservoir were sourced from the public streamflow database (CEDEX, 2024) (these were resampled from daily to monthly timesteps), along with monthly precipitation and mean temperature. This data was split into training (initial 80% of the timeseries) and testing data (final 20% of the timeseries).
- **Feature engineering:** lagged features (i.e. feature values at an earlier timestep) were created from the monthly time series (reservoir storage up to a year, meteorological variables for the preceding 6 months), and averaged features (mean reservoir storage for a given month, mean precipitation and temperature for the previous 3, 6 and 12 months). Month of the year was converted to a cyclic feature (using sine and cosine functions).
- **Model selection:** a range of ensemble-tree models were applied, with the best performing models for each reservoir taken forward. The models trialled were: Ada Boost Regressor; Bagging Regressor; Random Forest Regressor; Extra Trees Regressor; and Gradient Boosting Regressor, all implemented through the *Scikit-learn* Python package (Pedregosa F. et al., 2011). For each reservoir, the Extra Trees Regressor (ET) was selected based on training and testing scores.
- **Feature selection:** for each model, the most significant features were taken forward, as determined using the feature importance property of the model. Various thresholds for importance were explored, and the one that gave the greatest improvement in model performance was chosen to determine which features were "significant" enough to keep.
- **Hyperparameter tuning:** a random search method was used to select the hyperparameters which give the best model performance while minimising overfitting. Hyperparameters that were varied include: number of estimators (trees); maximum number of features considered at each split; maximum depth of trees; minimum number of samples per split; minimum number of samples per leaf; and whether or not bootstrapping is used.
- **Prediction and evaluation:** the model was used to predict reservoir storage at 1 and 3 month lead times, using a recursive approach for multi-step prediction. An ARIMA model was used as a baseline for model comparison, built using the same selected features as the ET models using the *statsmodels* Python module (Seabold & Perktold, 2010). The structure is ARIMA(1,0,1)$(2,1,0)_{12}$ as determined by the auto_*arima* function in the *pmdarima* Python library.

The following steps are employed to build the multi-reservoir model:

- **Data selection:** monthly reservoir storage, precipitation, and mean temperature were sourced as in the individual models, with reservoir characteristics and additional timeseries from the UK (NRFA, 2022). Catchment characteristics were sourced from the Caravan and CAMELS datasets (Kratzert et al., 2023; Delaigue et al., 2024). The timeseries data for each reservoir do not have identical start and end dates, so the data was split into training and testing sets by taking the initial 75% of the total time period as training data and the final 25% as testing data (i.e. all data points prior to May 2010 are in the training set, and all points after May 2010 are in the test set). This was done to ensure no data leakage from the testing to the training set. Since there is more data available later in the time period, this leads to a balance

of 67% of data points in the training data (and 33% in the testing). This split was chosen so that some reservoirs were solely in the testing period which allowed us to evaluate model performance on entirely unseen reservoirs.

- **Feature engineering:** lagged and averaged variables for reservoir storage and meteorological variables were used as in the individual models, but with storage converted to a percentage of total capacity so that the target variable is more consistent across the different reservoirs. Reservoir and catchment characteristics are also used, as detailed in Table 4. Categorical variables were One-Hot encoded.
- **Model selection:** Multiple model types were not explored for the global model to reduce computational effort, instead the Extra Trees Regressor was chosen based on the results of the model selection process for the individual reservoirs.
- **Feature selection:** a process of feature elimination was used here. Each feature was removed in turn and the model retrained, then the model performance was compared to the performance with all the features included for each reservoir. This demonstrated that, for many features, model skill across the reservoirs improved if that feature was removed (change in model skill was calculated as $\Delta skill = \frac{NSE_{drop} - NSE_{all}}{NSE_{opt} - NSE_{all}}$, where NSE is the Nash-Sutcliffe efficiency metric, for model simulations with one feature removed (drop), all the features (all), and optimum NSE (opt) which is equal to 1). Using different thresholds for median change in model skill, features were removed until optimum median model performance across all the reservoirs was reached.
- **Hyperparameter tuning:** a random search method was used to produce hyperparameter sets (including all the hyperparameters tuned for the individual models), the hyperparameters that gave the best median model performance across all the reservoirs were selected.
- **Prediction and evaluation:** simulations of reservoir storage at 1 and 3 months lead time were evaluated on the training data using the NSE metric, and compared to the individual models where available (for this, the evaluation period was cropped to match that of the individual models).