Author response to Referee #2 comments:

The authors have responded some of my concern while it is still not clear to me what changes and improvement have been made during this round. The responses are either vaguely stated in the response letter or could not find the revisions in the updated manuscript. Additionally, the differences between the observations and ERA5 data are still not well explained. Most figures still need significant improvement. A much clear point-to-point response letter and revised manuscript should be added.

Dear Reviewer,

Thank you for your additional comments. The EGU journals follow an interactive public peer-review process. During the current discussion stage, we are expected to provide a *revision plan* rather than submit a fully revised manuscript. The actual revision will be conducted after the discussion period closes (05 December 2025). This is why the manuscript has not yet been updated in the system. A comprehensive revision report will be submitted at that time.

Nevertheless, we have already addressed your comments. Therefore, a more detailed, point-by-point response with clarifications regarding the ERA5 data and observational datasets is provided below. The updated figures are also attached at the end of response letter.

Best regards,

Songjun Wu, on behalf of all co-authors

Reviewing of the manuscript open for discussion 'EcoTWIN 1.0: A Fully Distributed Tracer-Aided Ecohydrological Model Tracking Water, Isotopes, and Nutrients' by Songjun Wu et al. submitted to Geoscientific Model Development (Manuscript ID: egusphere-2025-3941).

The authors have developed/improved and ecohydrological model with advanced traits, tracking water, isotopic and nutrient fluxes. The reviewer considers that this model development work is worthwhile and beneficial to the modeling community and fall within well with the journal topic. There are a couple of issues that the reviewer suggests for the authors to refer, and they are listed in detail as follows

** We appreciate the positive evaluation by the reviewer. All comments have been addressed accordingly.

Almost all figures need to be replotted. Here I give some examples. Fig. 1: The texts are too small to read, and the subtitles need add some explanations for the figure, so that the reader is easy to follow it; Fig. 2: the coordinates need to be added, and some important landmarks or rivers information should better be added; the texts are to small to be identified; Fig. 3: the color for the color should be more identifiable, rather than orange and blue two colors; Fig. 4: the time series figures give no information to the reviewer; similar issues also for Fig. 5-7, and the reviewer will NOT list all issues for the figures. The authors should check them carefully and revise them all.

** Thank you for suggestions. We have revised the figures accordingly. The updated figures are provided at the end of this response letter.

Fig. 1: Irrelevant text has been removed from the inset plot. We have also increased the size of remaining text in inset plot and improved the captions.

Fig. 2: The coordinates have been added, and text size has been increased. However, we hesitated to add river information for two reasons: (1) the figure is already crowded, and (2) adding only major rivers will potentially mislead readers on the actual dense network in our distributed modelling.

Fig. 3: We have adjusted the scale of color code to make data points more identifiable (KGE ranges from [0, 0.5] to [0, 0.7].

Fig. 4: We respectfully argue that the time series gives readers intuitive impressions of the model's performance. But we have reduced the number of subplots for simplicity.

Fig. 5-7: The spatial plots show model performance in each grid cell compared to three remote sensing products. These are important validations for the ability of a hydrological model to reproduce uncalibrated hydrological fluxes, or namely the physical consistency. Therefore, we keep the three figures. The ambiguity may originate from the misleading color code of the spatial map, which has been changed to "*cool-warm*" to differentiate from the inset subplots.

The structure of this manuscript may be adjusted. Part 4.1 and 4.2 may be moved under the section of 3 Model calibration and validation? This should be one section to discuss the model advantages of accuracy compared with previous models? For example, providing specific skill metrics values.

** Yes, we agree that merging Section 4.1 and 4.2 to model calibration and validation section fits better to main streamline. This is implemented (new section 3.4 and 3.5). Meanwhile, the water age section 4.3 is now independent result section 4.

The updated outlines are:

…
3. Model calibration and validation
- *3.1 Model setup*
- *3.2 Model calibration*
- *3.3 Model validation*
- *3.4 Calibration performance*
- *3.5 Validation performance*

4. Water age simulations and its link to water quality
5. Discussion
…

Provide some explanations for the reason why these 17 catchments are selected for this study.

** The watersheds were selected due to data availability, particularly in-stream isotopes and nitrate with sparse distribution across Europe. Clarification has been added Section 3.

To ensure model generality, 17 catchments were selected for calibration and validation depending on the data availability (particularly stream stable water isotopes and nitrate), which span a wide range of characteristics in geography, climate, and anthropogenic managements (Figure 2 and Table 1).

What is the difference between model calibration and validation? The reviewer could not quite understand herein, it seems the difference is defined by the different variables that are choose for the model-to-data comparison. Please explain it.

** Like most hydrological and water quality models, we calibrated EcoTWIN with spatially distributed point observations, including discharge, in-stream isotopes, and nitrate. However, this does not guarantee the physical consistency of uncalibrated internal fluxes (e.g., snow melt/accumulation, evapotranspiration, percolation, etc.). Therefore, additional validation was applied for those uncalibrated states/fluxes. Three remote sensing products were used to test or informally validate EcoTWIN's ability of reproducing snow depth, evapotranspiration, and total water storage without direct calibration (against these variables). This is now clearly clarified in method section 3.

A robust model application should not only reproduce observed variables through calibration but also yield realistic estimates of internal states and fluxes that are not included in the calibration process. This is essential to avoid situations where inaccurate process representations produce deceptively good results through error compensation. Therefore, we evaluate EcoTWIN from both perspectives. First, we

assess the model's ability to reproduce observations via calibration (methods and results in Sections 3.2 and 3.4). Then, we examine the model's capacity to simulate uncalibrated internal states and fluxes by comparing the simulated snow depth, evapotranspiration, and total water storage with corresponding remote-sensing products (methods and results in Sections 3.3 and 3.5).

More skill metrics should be defined and used for model performance. For example, Root Mean Square Error, Correlation Coefficient, Mean Value Difference etc. to evaluate the model performance. Then, give the statistics and compare them with previous other models.

** We agree that the benefits of including more metrics. A new Table 2 has been added with statistics of different metrics (Kling–Gupta efficiency, root mean square error, Pearson correlation coefficient, percent bias). A brief comparison with literature has also been added to section 3.4.

The calibration was validated using Kling-Gupta efficiency (KGE), Root Mean Square Error (RMSE), Pearson Correlation Coefficient (Coefficient), and Percent bias (Pbias) (Table 2).

Discharge simulation: Such performance is comparable to or better than previous continental calibration of hydrological models (e.g., ParFlow, Naz et al., 2023; E-HYPE, Donnelly et al., 2016).

Isotope calibration: Such simulation deviation due to the uncertainty in data and boundary initialisation is often reported in previous calibration (Smith et al., 2021).

Nitrogen calibration: In general, the model produces comparable performances to existing nitrogen modelling at the catchment (Wu et al., 2022, 2025b; Yang et al., 2018) and continental scales (Jones et al., 2023; Mikayilov et al., 2015).

Table 2. The calibration performance of discharge (Q), in-stream isotopes (^{18}O, Iso), and nitrate (NO^3-N). Evaluation metrics include Kling-Gupta efficiency (KGE, -), Root Mean Square Error (RMSE; m3/s, ‰, and mg/L for discharge, isotopes, and nitrate, respectively), Pearson Correlation Coefficient (Coefficient; -), and Percent bias (Pbias; %).

Metric	Min	Max	Mean	Median
KGE (Q)	0.14	0.89	0.65	0.69
KGE (Iso)	-0.03	0.86	0.45	0.48
KGE (NO^3-N)	-0.36	0.72	0.42	0.44
Correlation (Q)	0.49	0.92	0.79	0.81
Correlation (Iso)	0.14	0.87	0.51	0.54
Correlation (NO^3-N)	-0.26	0.86	0.55	0.6
RMSE (Q)	3.99	677.08	123.02	68.51
RMSE (Iso)	0.31	1.51	0.72	0.73
RMSE (NO^3-N)	0.02	2.82	0.83	0.57
Pbias (Q)	0.52	79.88	17.44	9.53
Pbias (Iso)	-11.28	-0.07	-4.3	-4.42
Pbias (NO^3-N)	0.18	49.25	15.52	10.89

Some small mistakes or errors. Table 1, the table better not crossing two pages. Line 378, add the equation number and the meaning of it (e.g., what does l mean in the left of the equation). Be consistent use KEG or KGE, and define it and explain it (e.g., what are the value ranges, and the corresponding performance, excellent, good, normal, poor etc.)

** Thank you for suggestions. All tables have been moved to an independent section to avoid page crossing. The equation has been labeled with explicit description of L (likelihood). KEG is the typo of KGE (Kling-Gupta efficiency). This has been corrected through text.

Line 451: ERA5 reanalysis products equal to observations? Please double check it.

** Yes, the snow depth simulated in model and in ERA5 reanalysis is commensurable, but there are still uncertainties and differences between ERA5 products and real observations. We now use "ERA5 records" to avoid ambiguity.

Finally, the simulated snow depth was compared to the daily snow depth in ERA5 reanalysis products (ERA5 post-processed daily statistics on single levels; 10.24381/cds.4991cf48). Results in Figure 7 show a good agreement between simulations and ERA5 records in most regions with r2 > 0.5, though degradation was found in a few catchments.

Line 531-532: 'However, it increases severely increase the ……' Delete one of the 'increases'.

** Thank you for pointing out. This has been corrected.

Line 694: the first letter of the word 'mediterranean' should be initialized.

** Mediterranean is now capitalised through text.

Updated Figures:

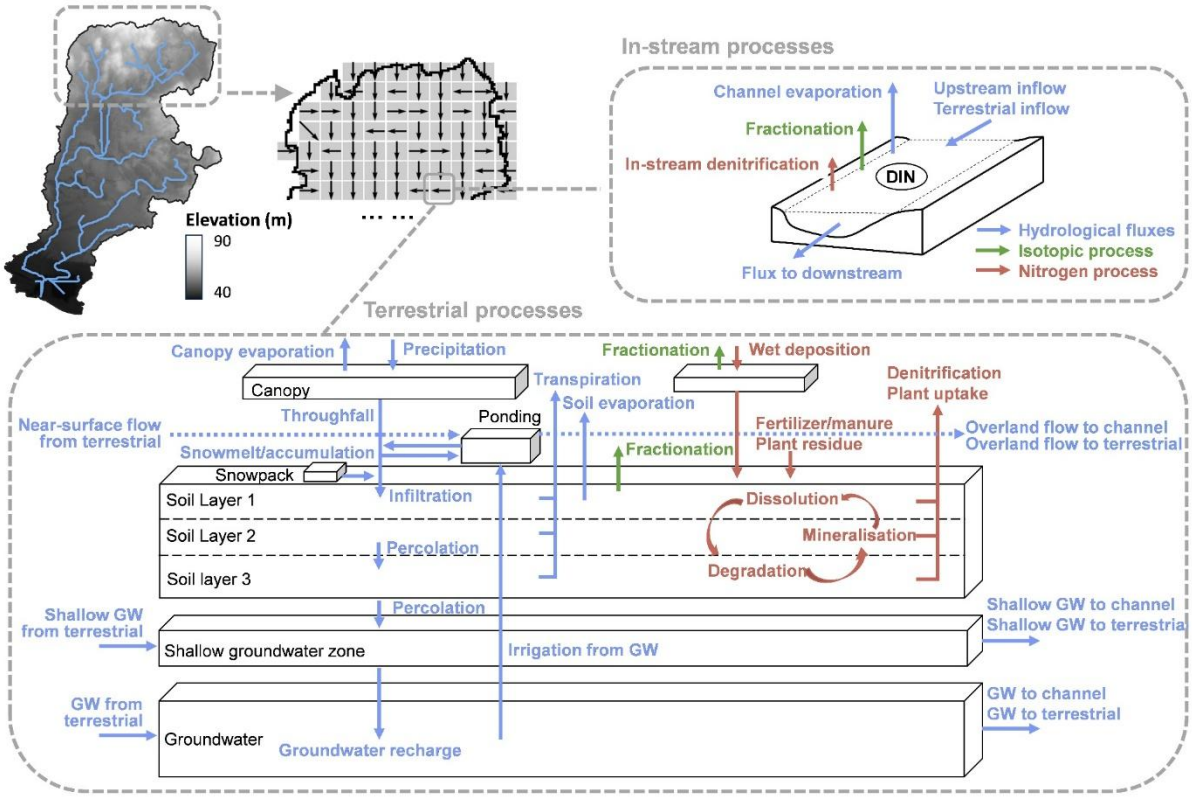

Figure 1. Model structure of EcoTWIN. As a distributed model, EcoTWIN disentangles the spatial domain into grid cells. In each grid cell, hydrological, isotopic, and nitrogen processes were simulated in canopy, snow, soils, shallow groundwater, groundwater, and channel.

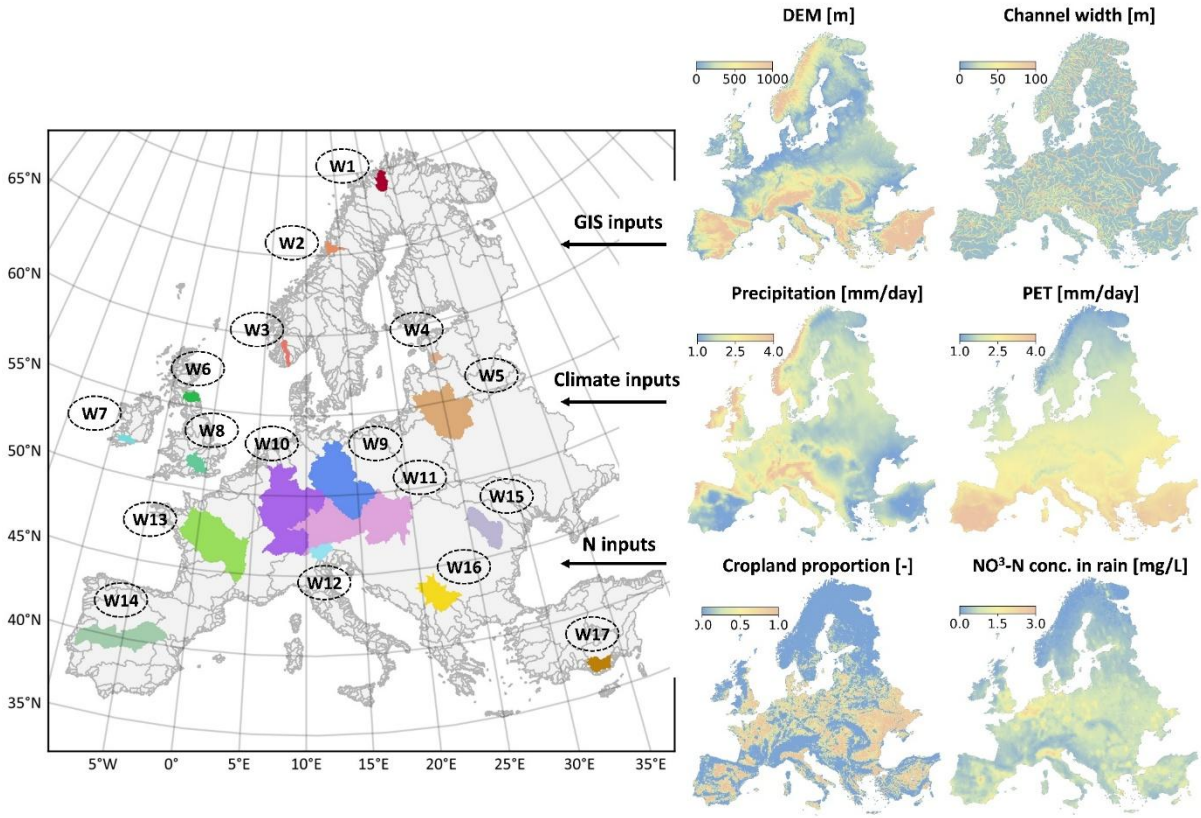

Figure 2. The selected catchments for model validation and an overview of key inputs.

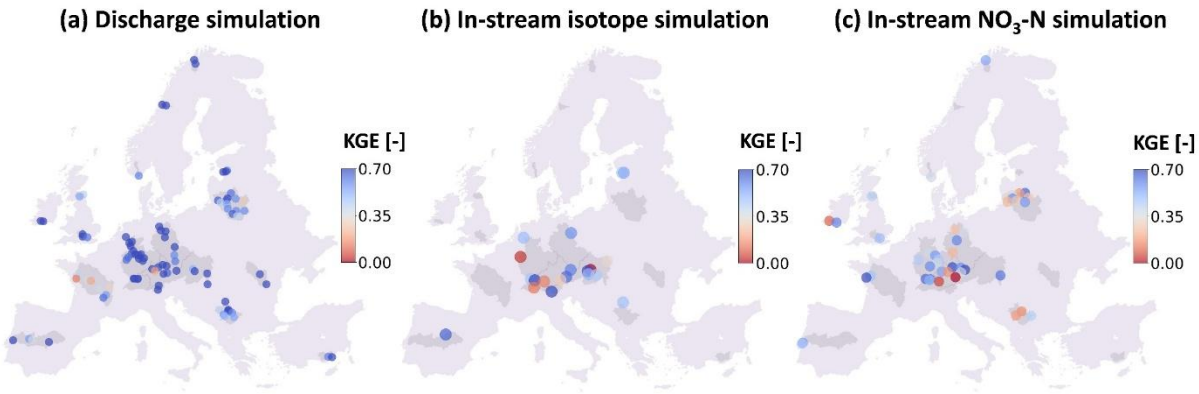

Figure 3. The simulation performance of discharge, in-stream isotope, and in-stream NO_3-N.

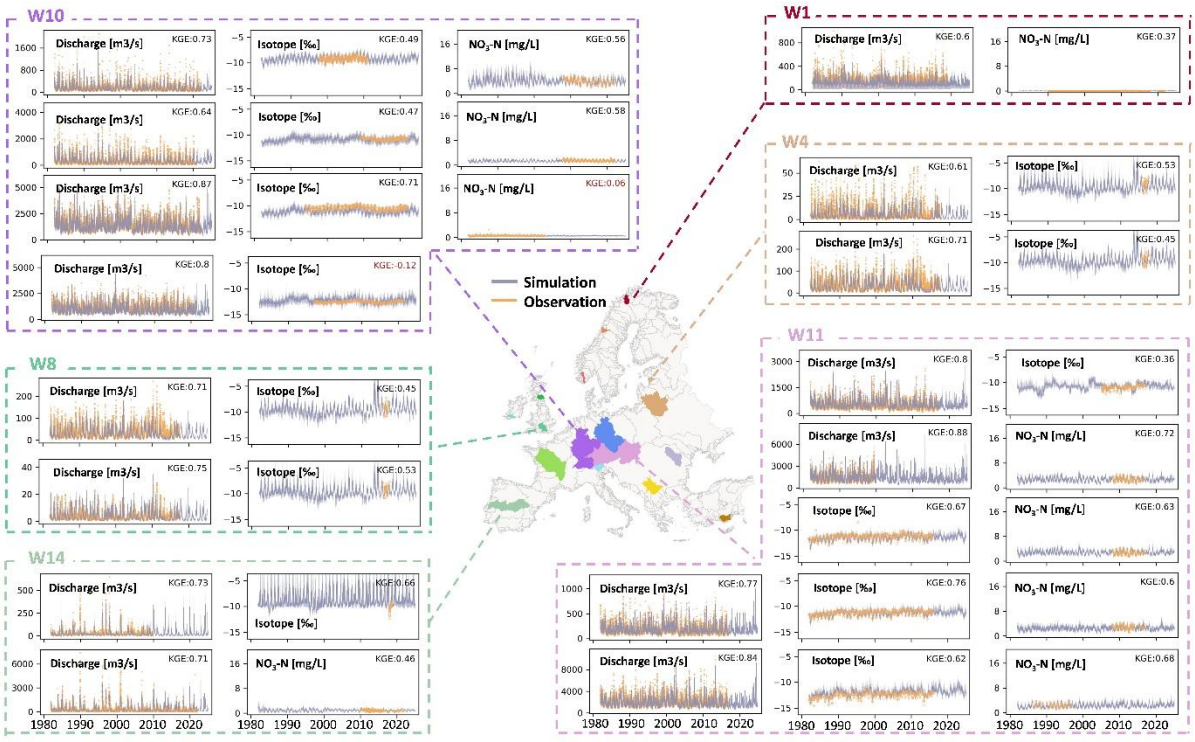

Figure 4. The simulated (blue) and observed (orange) time series of discharge, isotopes, and NO_3-N at representative gauges. Note that the sites with relatively poor performance (KGE < 0.2) were particularly shown for model diagnosis.

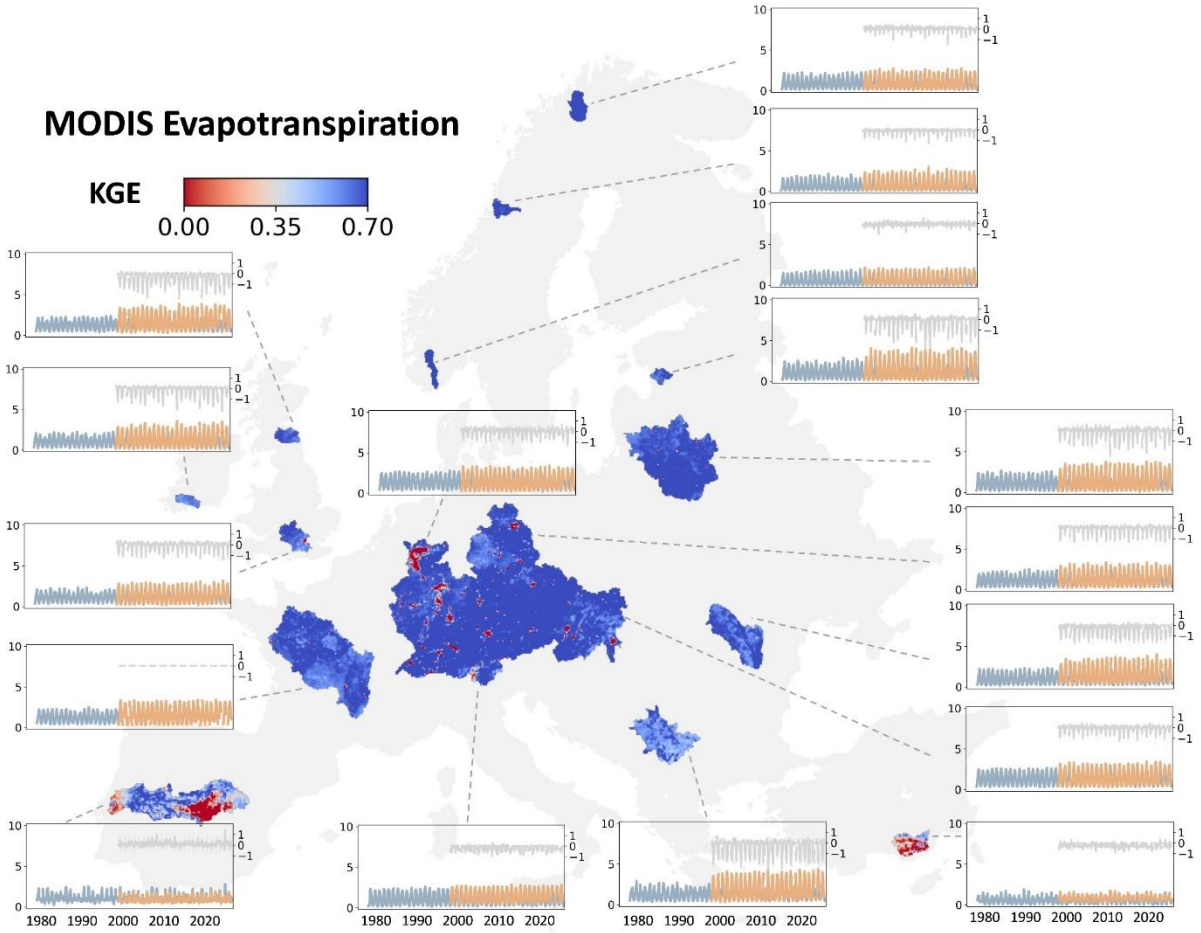

Figure 5. The grid-to-grid comparison between simulated evapotranspiration and MODIS evapotranspiration shown in KGE. The time series in inset subplots show the monthly dynamics of simulated (blue) and observed (orange) values averaged from all grid cells in the watershed, as well as their deviations (grey).

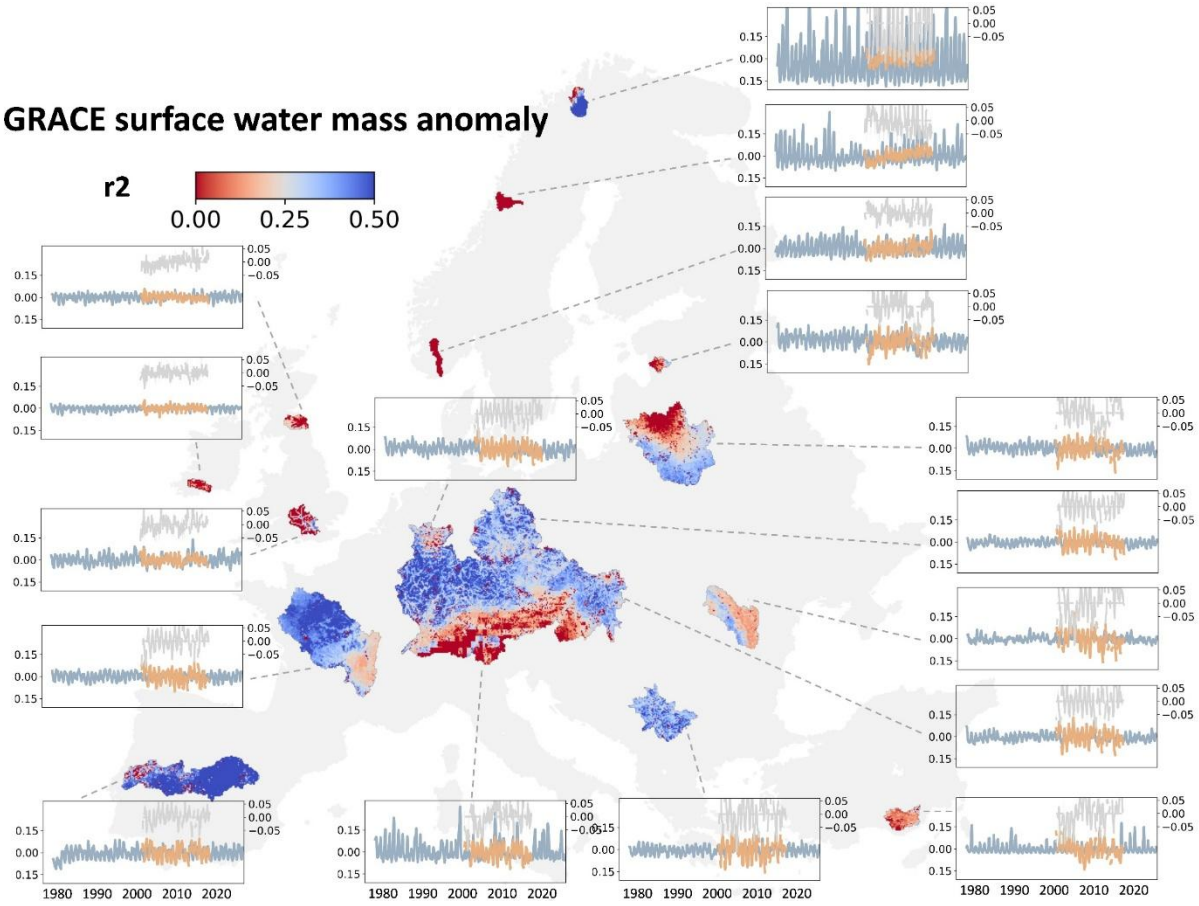

Figure 6. The grid-to-grid comparison between simulated water storage anomaly and GRACE surface water mass anomaly. The time series show the monthly dynamics of simulated (blue) and observed (orange) values averaged from all grid cells in the watershed, as well as their deviations (grey).

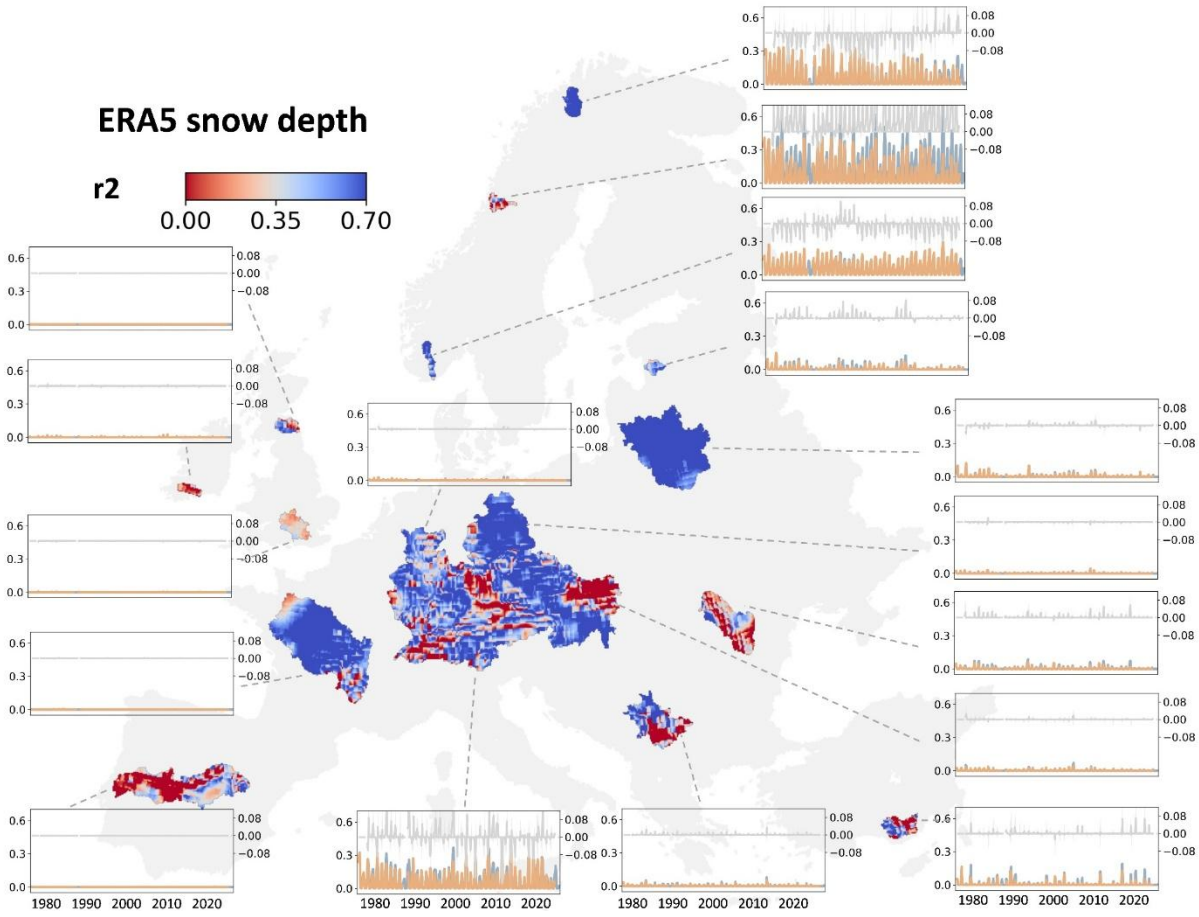

Figure 7. The grid-to-grid comparison between simulated snow depth and ERA5 snow depth. The time series show the monthly dynamics of simulated (blue) and observed (orange) values averaged from all grid cells in the watershed, as well as their deviations (grey).