Technical Appendix to EGU26-77: Multisite and Multivariate Calibration of the SWAT+ Model in Galicia-Costa, NW Spain

Darine Saad[1]*, Carolina Acuña Alonso[1], Xana Álvarez Bermudez[1]

[1]*Universidade de Vigo, HydroForestry Geomodeling Research Group, 36005 Pontevedra, Spain.*

*Contact: darine.saad@uvigo.gal

ABSTRACT

Calibration and validation are fundamental steps in hydrological modeling, ensuring the accuracy and reliability of model predictions for effective management of freshwater resources. Traditional calibration approaches rely solely on streamflow, although other hydrological variables (soil moisture, evapotranspiration) have also proved useful. One of the most widely used hydrological models is the Soil and Water Assessment Tool (SWAT), which simulates spatial and temporal variations in watershed processes such as the water balance, streamflow routing, and the transport of nutrients and sediments. In this study, SWAT+, a revised version of the SWAT model, was applied to the Ulla River basin within the Galicia-Costa Hydrographic Demarcation in Spain to evaluate the performance of three calibration strategies: (1) single-variable calibration using streamflow (SC-Q), (2) single-variable calibration using evapotranspiration (SC-ET), and (3) multivariate calibration, integrating both streamflow and evapotranspiration (MC-QET). Multi-site calibration and validation were performed using the Sequential Uncertainty Fitting Algorithm (SUFI-2), with the Nash-Sutcliffe efficiency (NSE) index as the objective function and NSE ≥ 0.60 defined as the behavioral threshold. Observed streamflow data was obtained from three river gauging stations distributed along the river network (one downstream and two upstream). Ground-truth evapotranspiration (ET) data were estimated via triple collocation analysis combining three independent datasets (remote sensing-based, land surface model output, and reanalysis product). Results revealed that for streamflow, the MC-QET calibration scheme yielded the best performance at the downstream validation site (NSE = 0.82, PBIAS = -4.51), whereas SC-Q achieved superior results at the upstream stations (NSE = 0.82-0.86 and PBIAS = +6.35 – +12.72). Meanwhile, SC-ET performed the worst for streamflow overall, although model performance was still acceptable (NSE = 0.70 – 0.75). For evapotranspiration, both SC-ET (NSE = 0.89, PBIAS = +6.56) and MC-QET (NSE = 0.90, PBIAS = +6.24) clearly outperformed SC-Q (NSE = 0.66, PBIAS = +22.97). These findings suggest that while streamflow- or ET-only calibration can optimize the targeted variable, incorporating multiple hydrological variables during model calibration improves the overall representation of

watershed processes and the water balance. However, the acceptable performance of the ET-only calibration highlights that this calibration scheme can still serve as a valid alternative in data-scarce regions where streamflow observations are limited or inconsistent. Furthermore, this study demonstrates the reliability of triple collocation analysis in improving ET estimates by reducing uncertainty among independent data sources. In conclusion, integrating multivariate calibration strategies in SWAT+ significantly enhances spatial transferability, ensures physically realistic model outputs, and improves overall prediction reliability. At the same time, ET-based calibration alone remains a practical and defensible option for data-limited watersheds, demonstrating the growing potential of remote sensing–driven hydrological modeling for comprehensive and resilient water resource assessment.

Citation:

Saad, D., Acuña Alonso, C., & Álvarez Bermúdez, X. (2026, May 3–8). *Multisite and multivariate calibration of the SWAT+ model in Galicia-Costa (NW Spain)* [Poster abstract]. European Geosciences Union General Assembly (EGU26), Vienna, Austria. https://doi.org/10.5194/egusphere-egu26-77.

QR Code:

Contents

List of Figures

List of Tables

List of Annexes

1. Introduction

This supplementary document serves as a **technical appendix to Abstract EGU26-77 presented at the European Geosciences Union General Assembly 2026 (EGU26)**. It provides a detailed description of the datasets, processing procedures, and methodological approaches used in the analysis, as well as the complete set of results generated in this study. The aim of this document is to ensure transparency, clarity, and reproducibility of the methods and findings.

2. Study area

The study area focuses on the Ulla River catchment, located in the Galicia-Costa Hydrographic Demarcation (GCHD) in the autonomous community of Galicia, NW Spain. The Ulla River is the second largest river in the region with a catchment area of 2,803 km^2. It originates from Fonte de Ulloa in the municipality of Monterosso at around 600 m altitude and flows across the middle of Galicia with a watercourse bed length of 142 km in a southwesterly direction, crossing 19 counties. Finally, it discharges into the Atlantic Ocean through the Ria de Arousa at a mean annual flow rate of 79.3 m^3/s (Farinango et al., 2023; Oliveira et al., 2020).

This study specifically considers the fluvial section of the basin, with the outlet defined upstream of the town of Padrón (**Figure 1**). This delineated catchment area comprises 2,400 km^2, representing approximately 85% of the total basin, and is hereafter referred to as the "Ulla Basin". The remaining area primarily comprises the estuarine transition zone and was excluded to focus on the non-tidal freshwater dynamics.

Land use in the Ulla Basin is predominantly comprised of natural and semi-natural land, covering about 66% of the total area. These lands consist mainly of forests and forest plantations, along with shrubland, grassland, woodland, and wetlands. Agricultural activities, including croplands and pastures, are the second most common land use type, occupying 31% of the basin. The remaining 3% are primarily urban settlements (MITECO, 2021). The dominant soil units in the study area are mainly Umbric Leptosols and Humic Cambisols, occupying 30% and 70% of the basin area, respectively (FAO & IIASA, 2023)

Three reservoirs are located along the Ulla River: Portodemouros, Bandariz and Touro. They are mainly used for energy production (hydroelectricity) and flood control. Portodemouros is the largest, with a capacity of 297 hm^3, while Bandariz and Touro are smaller, holding 2.74 hm^3 and 3.78 hm^3, respectively (Augas de Galicia, 2025).

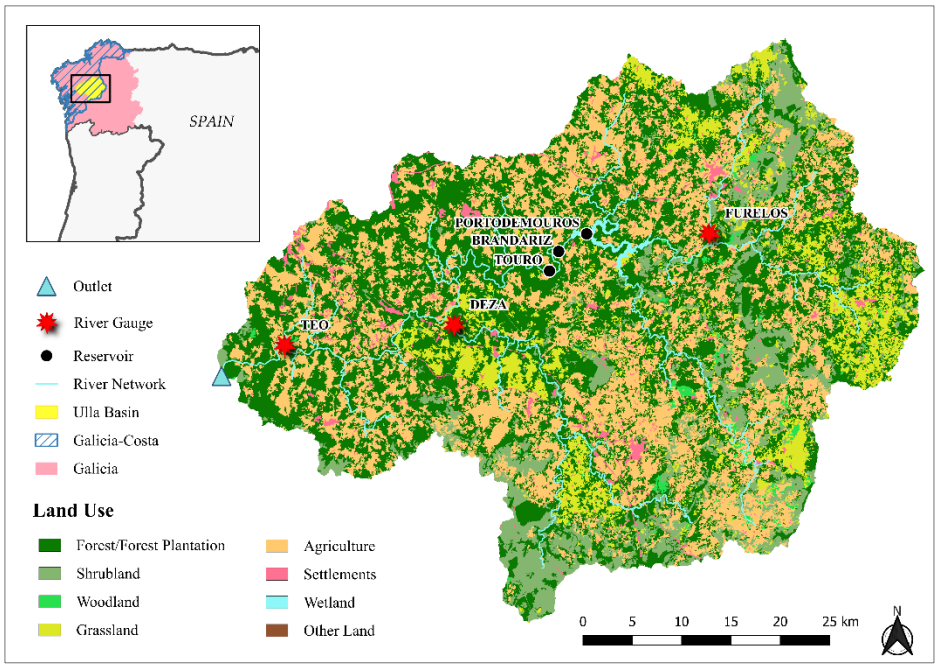


Figure 1. *Map of the Ulla Basin study area with dominant land uses.*

3. The Soil & Water Assessment Tool (SWAT) and SWAT+

The Soil and Water Assessment Tool or SWAT (Arnold et al., 1998) is a semi-distributed, watershed-scale, process-based ecohydrological model that is widely used to simulate water quality and quantity. It was developed and designed by the Agricultural Research Service (ARS) of the United States Department of Agriculture (USDA) to model spatial and temporal variations in watershed processes such as the water balance, streamflow routing, and the transport of nutrients and sediments. The SWAT model is commonly used to assess the impact of climate change, land use and land management practices on water supply, nutrient loads, and sediments transport within river basins and watersheds of different scales (Abbas et al., 2024).

SWAT+ (Bieger et al., 2017) is a revised version of the SWAT model. While still based on the same equations, this updated version provides more flexibility in water, nutrient, and sediment routing between spatial objects in the watershed as well as in defining management schedules. The SWAT+ model operates based on hydrological response units (HRUs), consisting of areas within each subbasin that share the same land use, soil type, and slope class. A new feature of SWAT+ is also the landscape units (LSUs), which are subdivisions of the subbasins that separate upland processes from floodplains and are collections of HRUs (Bieger et al., 2017). In this sense, a watershed in the SWAT+ model is delineated into subbasins based on a DEM layer and a user-defined stream threshold, and each subbasin is in turn divided into LSUs and HRUs according to channel thresholds, as well as land use, soil type, and slope class distribution. The water balance equation used by SWAT+ is shown below.

$$SW_t = SW_o + \sum \left(R_{day} - Q_{surf} - E_\alpha - W_{seep} - Q_{gw} \right)$$

Where SW_t is the final and SW_o the initial water content at a given time (days), Q_{surf} the surface runoff generated from precipitation on a given rainy day (R_{day}), E_α the evapotranspiration, W_{seep} the percolation, and Q_{gw} the amount of baseflow on a given day, with all components expressed in mm (Castellanos-Osorio et al., 2023).

3.1. Datasets and software description

3.1.1. Land cover and land use

A land cover and land use map for the study area was obtained from the "*Foto Fija del Mapa Forestal de España*" (Still Photo of the Forest Map of Spain) for the year 2021, with a resolution of 1:50,000 and 1:25,000. The data was obtained from MITECO (*Ministerio para la Transición Ecológica y el Reto Demográfico*), the Ministry for the Ecological Transition and the Demographic Challenge in Spain (MITECO, 2021).

The map was reclassified and detailed land use classes were assigned to match the classification scheme used by the SWAT+ model. Where available, information on the dominant species in each land use type was also incorporated into the reclassification. If a land use type was not directly available in the database, the closest matching category from the SWAT+ classification was assigned instead. **Table 1** describes the land use types identified in the Ulla Basin with their corresponding SWAT+ land use codes.

Table 1. *Land use classes in the Ulla Basin and their corresponding SWAT+ code.*

Land use type	SWAT+ code
Agricultural land	agrc
Barren or sparsely vegetated land	bsvg
Cropland/grassland mosaic	crgr
Cropland/woodland mosaic	crwo
Deciduous broadleaf forest	fodb
Deciduous forest - temperate oceanic	frsd_teof
Eucalyptus plantation	euca
Evergreen forest - temperate oceanic	frse_teof
Evergreen needleleaf forest	foen
Forested wetland	wetf
Grassland	gras
Industrial area	uidu
Mixed forest	frst
Mixed forest - temperate oceanic	frst_teof
Mixed grassland/shrubland	migs
Oak	oak

Pasture	past
Pine	pine
Poplar	popl
Shrubland	shrb
Transportation	utrn
Urban area	urbn
Water bodies	watr

3.1.2. Soil data

A soil map covering the study area was obtained from FAO's Harmonized World Soil Database v2.0, with a resolution of 1 km and 30 arc-second (FAO & IIASA, 2023). When multiple soils were found within the same mapping unit, only the dominant soil was kept. Two main soil units were identified in the study area: Umbric Leptosols and Humic Cambisols (**Figure 2**). Properties for each soil type were obtained from the FAO database where available, as well as from consulting the literature (Guarracino, 2007; Post et al., 2000; Williams, 1995).

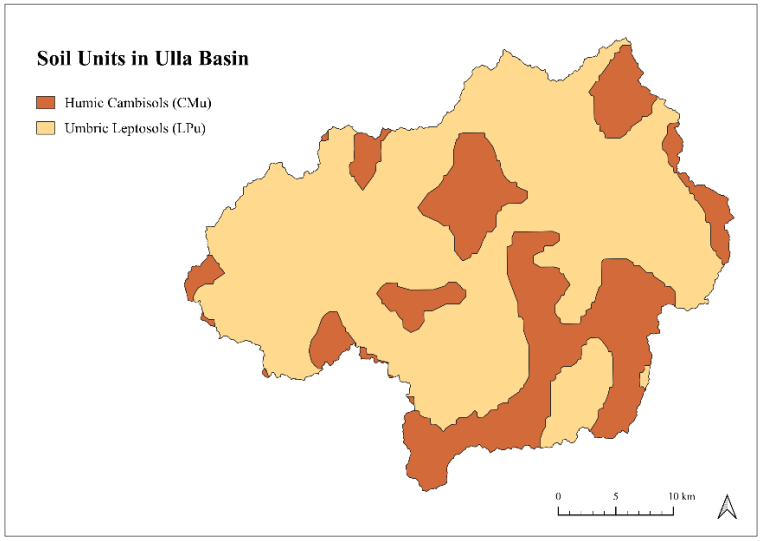


Figure 2. *Map of the dominant soil units in the Ulla River basin.*

3.1.3. DEM

A Digital Terrain Model (DTM) with a resolution of 25 m was obtained from the National Center for Geographic Information of Spain (CNIG, 2025). The DTM was preprocessed to fill depressions based on L. Wang & Liu (2006) and prepare it for watershed delineation.

3.1.4. Meteorological data

Weather observations from 17 relevant weather stations (**Table 2**) distributed across the study region were obtained from MeteoGalicia, the meteorological agency of Galicia, Spain (MeteoGalicia, 2025). Data marked as "non-validated", "suspicious", or "incorrect" were all treated as missing data. Daily observations were obtained for the following variables, as required by the SWAT+ model: precipitation, maximum and minimum temperature, wind speed, relative humidity, and solar radiation.

Table 2. *Name, coordinates, and altitude of the weather stations in the Ulla Basin study area.*

Station name	ID	Latitude (°)	Longitude (°)	Elevation (m)
Arzua	10144	42.931965	-8.174690	362
Sergude	10095	42.822826	-8.461246	231
Melide	10089	42.907383	-7.982653	477
Santiago-EOAS	10124	42.875960	-8.559434	255
Santiago-San Lazaro	50500	42.886590	-8.521138	305
Pazo de Galegos	19003	42.784780	-8.434715	225
Alto do Faro	10202	42.669987	-7.887967	842
Corno do Boi	10101	43.037400	-7.892649	731
Serra Vacaloura	10170	42.816715	-7.750347	780
Xesteiras	10133	42.675583	-8.586177	715
Forcarei	10163	42.610943	-8.370965	674
Mouriscade	10061	42.614384	-8.137295	500
Lalin	10158	42.656265	-8.107154	601
Serra do Faro	10122	42.579758	-7.933075	991
Camanzo	19050	42.779110	-8.318691	211
Amiudal	10109	42.414585	-8.238242	553
Costa	19011	42.796500	-8.716000	111

When weather observations are missing or unavailable, SWAT+ can simulate values using the built-in WXGEN weather generator (Sharpley & Williams, 1990). A weather generator database for Peninsular Spain is provided on the SWAT website (https://swat.tamu.edu/data/spain/). This dataset contains monthly weather data from 1979 to 2013, derived from rainfall and temperature data provided by the Spanish National Meteorological Service (AEMET) with a resolution of 5 km, along with relative humidity, solar radiation, and wind speed obtained from the Climate Forecast System Reanalysis (CFSR) with a resolution of 38 km (Senent-Aparicio et al., 2021).

3.1.5. Observed streamflow data

Records of streamflow observations were obtained from the download portal of MeteoGalicia (https://servizos.meteogalicia.gal/mgafos/estacions/estacions.action). This data is sourced from records of the gauging stations operated by Augas de Galicia, the

hydraulic administration of the Galician community. Observed streamflow data for 3 sites along the Ulla River (two upstream and one downstream; see **Figure 1**) were used in model calibration and validation, measured from 3 river gauging stations: Furelos and Deza (upstream), Teo (downstream). A multi-site calibration approach was adopted as previous studies have demonstrated it to be the most effective choice (Aragaw & Mishra, 2022; Makumbura et al., 2022; Singh & Saravanan, 2022).

3.1.6. Software

Spatial analysis was performed in QGIS (version 3.34.14), and data analysis using R (version 4.4.2) in RStudio.

3.2. SWAT+ model setup

The model set-up in SWAT+ (version 3.0.8) consisted of three major steps: (1) preparation of input data, (2) delineation of the watershed, and (3) definition of dominant HRUs. Input data consisted of the DTM layer, land use map, and soil map, all in raster format. The land use and soil maps were resampled to match the resolution of the DTM layer (25 m). A predefined river network was obtained from SERGAS (https://www.sergas.es/Saude-publica/GIS-Demarcacion-hidrografica-Galicia-Costa?idioma=es), and was overlayed on the DTM to be used as "burn-in" for a more accurate definition of the watershed steam network.

Delineation of the Ulla River watershed and its subbasins were performed using QSWAT+ (version 3.0.3), the QGIS interface for the SWAT+ model. The DTM was first set up as the primary input, and an outlet point was defined, as well as the three reservoirs along the Ulla River. Then, streams and the channel network were delineated using the input predefined river network as reference. Finally, the watershed and corresponding subbasins were delineated. HRUs were determined based on the dominant land use, soil class and slope band found in each subbasin. A total of three slope bands were defined: 0–5%; 5–10%; and >10%. Short channels that occupy less than 5% of the subbasin area were merged, and HRUs were filtered to retain only those occupying at least 5% land use, 10% soil type and 10% slope class of a landscape unit in a way to reserve the spatial heterogeneity of the study area.

The Ulla River watershed was delineated into 25 subbasins, with a total of 1000 HRUs. Simulations were run from 01/01/2000 to 01/01/2025 using the SWAT+ Editor (version 3.0.8), with the first 5 years treated as a warm-up period to ensure that "initial" state variables (i.e., soil moisture, reservoir levels, etc.) are stabilized before actual simulation begins. All parameters were kept as default for the first iteration, only adjusting and fixing key reservoir parameters (**Table 3**).

Table 3. *Reservoir parameters adjusted in SWAT+ Editor (P.S. = principal spillway; E.S. = emergency spillway).*

Reservoir	Area at P.S. (ha)	Volume at P.S. (10^4 m^3)	Area at E.S. (ha)	Volume at E.S. (10^4 m^3)
Portodemouros	1076.03	29700	1237.43	34200
Bandariz	56.99	274	65.54	315.1
Touro	60.12	378	69.14	434.7

3.3. SWAT+ calibration and validation

While hydrological models are often calibrated and validated based on streamflow alone, introducing additional variables alongside streamflow (e.g., evapotranspiration) can improve model accuracy, especially in simulating the water balance (Franco & Bonumá, 2017; Koltsida & Kallioras, 2022). To investigate this further, three calibration schemes were evaluated:

- **SC-Q:** Single-variable calibration using streamflow, following traditional methods.
- **SC-ET:** Single-variable calibration using evapotranspiration (ET); the purpose of this scheme is to assess the reliability of ET-only calibration for watersheds where streamflow data is missing.
- **MC-QET:** Multi-variate calibration, integrating both streamflow and ET; the purpose of including this scheme is to investigate the advantages of multi-variate over single-variate calibration.

Sensitivity analysis, model calibration and uncertainty analysis were all performed using the RSWAT (version 4.01) package (Nguyen, 2024), a web-based interactive application in R for conducting parallel sensitivity analysis, calibration, and uncertainty analysis with the SWAT model and its modifications (e.g., SWAT+, SWAT-Carbon). RSWAT features a graphical user-friendly interface within R and supports open-source parallel processing for SWAT+. Developed using the *Shiny* package in R, it also provides a platform for researchers to discuss questions or inquiries (Nguyen et al., 2022). For these advantages, RSWAT was selected in this study for model calibration and optimization.

3.3.1. Model calibration

The model was calibrated:

- on a daily time step from 01/01/2005 to 31/12/2018 for streamflow at the upstream stations and from 01/10/2008 to 31/12/2018 at the downstream station,
- on a monthly time step from 01/2005 to 12/2015 for evapotranspiration at the LSU scale.

A global sensitivity analysis was first performed to assess the influence of model parameters on streamflow and ET using a multivariate regression approach with parameter sets generated by uniform Latin Hypercube Sampling (LHS), as presented in the following equation:

$$g = \alpha + \sum_{i=1}^{n} \beta_i b_i$$

Where g is the value of the objective function used, α is the regression constant, n is the number of parameters, β_i is the coefficient of the ith parameter, and b_i the relative significance of the ith parameter. The relative significance of each parameter b is then identified by conducting a t-test. The sensitivity of each parameter is evaluated according to the t-score and p-value, where a higher absolute value of the t-statistic and a lower p-value implicate higher parameter sensitivity (Nazari-Sharabian et al., 2020). Only parameters with a significant p-value ($p < 0.05$) are considered for the calibration and uncertainty analysis. More details on this sensitivity analysis approach can be found in Abbaspour (2015).

A total of 20 parameters having the greatest influence on streamflow and ET were selected for sensitivity analysis. These parameters were identified by reviewing previous studies and consulting the literature. In RSWAT, parameters can be altered in three ways: (1) replace ($x' = c$), (2) relative change ($x' = x + c*x$), and (3) absolute change ($x' = x + c$), where x is the original parameter value and x' is the new parameter value after a change "c" is applied (Nguyen et al., 2022). Description of the selected parameters, as well as the type and range of applied change are shown in **Table 4**.

The sensitivity analysis was performed with 2000 model runs for all calibration schemes.

Table 4. *SWAT+ parameters selected for sensitivity analysis with their description, type of change applied, and the minimum and maximum value of the change. R: replace; r: relative; a: absolute.*

Parameter	Description	Change	Min	Max
epco.hru	Plant uptake compensation factor	R	0.01	1
esco.hru	Soil evaporation compensation factor	R	0.01	1
cn2.hru	SCS curve number for soil moisture condition II	r	-0.3	0.3
perco.hru	Percolation coefficient (fraction)	r	-0.5	0.1
cn3_swf.hru	Curve number condition III for soil moisture factor	R	0	1
latq_co.hru	Lateral flow coefficient	R	0	1
awc.sol	Available water capacity of the soil layer (mm H_2O/mm)	a	-0.027	0.961
alpha.aqu	Baseflow recession factor (days)	R	0	1
flo_min.aqu	The minimum depth from the surface to the water table required for groundwater flow to occur (m)	R	0	50

surlag.bsn	The coefficient for surface runoff lag (days)	R	0.05	24
revap_co.aqu	Groundwater revap coefficient	R	0.02	0.2
revap_min.aqu	Threshold depth of shallow aquifer for revap to occur (m)	R	0	50
canmx.hru	The upper limit of canopy storage (mm/H_2O)	R	0	100
k.sol	Hydraulic conductivity (mm/hr)	r	-0.5	0.5
snofall_tmp.hru	Snowfall temperature (ºC)	R	-5	5
snomelt_tmp.hru	Snow-melt base temperature (ºC)	R	-5	5
evlai.bsn	Leaf area index at zero evaporation from water bodies	R	0	10
slope.hru	Land surface slope (m/m)	r	-0.3	0.3
chn.rte	Channel Manning's n	R	-0.01	0.3
chk.rte	Effective hydraulic conductivity of the channel alluvium (mm/hr)	R	-0.01	500

Automatic calibration and uncertainty analysis were performed using a similar approach to the sequential uncertainty fitting algorithm (SUFI-2), which simultaneously performs calibration and uncertainty analysis. A detailed description of this algorithm is found in Abbaspour et al. (2004) and Abbaspour (2015). In SUFI-2, parameter optimization is applied to parameter sets and not individual parameter values, in a way to explicitly identify any interaction between parameters. The intended purpose of the SUFI-2 calibration approach is not to find the best-fit parameter set, but instead to find the best range for each parameter that yields an acceptable model performance. The optimization process therefore aims to improve the defined objective function, which quantifies the discrepancy between observed and simulated values, all while accounting for uncertainties.

Prediction uncertainties are introduced from various sources, e.g., driving variables (i.e., weather data: rainfall, temperature, etc.), model parameters, model structure, and observed data (i.e., measured streamflow, ET), all of which are accounted for in the SUFI-2 algorithm. These uncertainties are quantified and evaluated by two factors, the *p*-factor and the *r*-factor. The *p*-factor represents the percentage of observed data that is bracketed by the 95% prediction uncertainty (95PPU) band, calculated at the 2.5th and 97.5th percentiles of the cumulative probability curve of the simulated variable. The *r*-factor represents the thickness of the 95PPU bracket, calculated as the ratio between the average width of the 95PPU band and the standard deviation of the observed variable. An optimal value for the *p*-factor is 1 (i.e., 100% of observed data are bracketed by the 95PPU), while an *r*-factor value closer to 0 is preferred, indicating minimal uncertainty in streamflow simulation (J. Guo & Su, 2019). Abbaspour et al. (2015) recommended working values of *p*-factor > 0.7 and *r*-factor < 1.5 for discharge simulation.

The automatic calibration was performed with one iteration of 5000 total model runs for each calibration scheme.

3.3.2. Model validation

The model was validated for the period of 2019-2025 for streamflow and 2016-2021 for ET, using a combination of performance and error metrics, regression analysis and graphical techniques, as suggested by Harmel et al. (2014).

Graphical techniques

Several guidelines and protocols for hydrological model evaluation suggest the use of graphical techniques to validate and assess model performance, e.g. D. N. Moriasi et al. (2007), Biondi et al. (2012), Harmel et al. (2014). In this project, the following plots were generated to evaluate model accuracy in simulating streamflow: time series plots, scatter plots, and flow duration curve (FDC) plots.

Performance and error metrics

Summary statistics, performance metrics, and error indices were calculated to evaluate model performance in simulating both streamflow and ET.

a. Summary statistics:

Comparison of summary statistics (i.e., mean, median, maximum, minimum, standard deviation) for observed and simulated values, as recommended by Harmel et al. (2014), should be part of the initial assessment of overall model performance, alongside goodness-of-fit metrics and error indices (described in the following sections). This comparison serves as a preliminary tool for early evaluation of model performance. In fact, if significant differences are found between these summary values, an accurate model simulation is unlikely, and further refinement is required before conducting a more detailed analysis.

b. Nash-Sutcliffe efficiency (NSE):

NSE is defined as a normalized statistical metric used for assessing the predictive accuracy of hydrological models. It represents the ratio between residual variance (or "noise") and variance in observed data (or "information"), providing a reliable measure of model performance (Nash & Sutcliffe, 1970). NSE is calculated as follows:

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2}\right]$$

Where O_i the observed data value, P_i is the *i*th simulated data value, and $\overline{O}$ is the observation mean, with n being the total number of observations.

NSE is a dimensionless metric that ranges from $-\infty$ to 1.0, with NSE = 1 indicating optimal model performance (prediction with 100% accuracy). Model performance is generally deemed acceptable for positive NSE values ($0.0 < NSE \leq 1.0$) and considered to make better predictions than by assuming the mean of observed data., while $NSE \leq 0.0$ means that the observational mean is a better predictor than simulated values.

c. Kling-Gupta efficiency (KGE):

KGE is a modified version of NSE developed by H. V. Gupta et al. (2009). This index decomposes the NSE index into three independent components of the hydrograph: linear correlation (r), bias (β), and relative variability between observed and simulated data (α). The KGE index is calculated by computing the Euclidian distance (ED) of these three components, as follows:

$$KGE = 1 - ED$$

$$ED = \sqrt{[s_r\,(r - 1)]^2 + [s_\alpha(\alpha - 1)]^2 + \left[s_\beta(\beta - 1)\right]^2}$$

Where ED is the Euclidian distance, r is the linear correlation coefficient between observed and simulated data, α is the variability ratio defined as the ratio between the standard deviations of simulated and observed values, β is the bias ratio defined as the ratio between means of simulated and observed values, while s_r, s_α, and s_β are scaling factors used to re-weight the relative importance of each component. Conventionally, it is assumed that all three components hold equal weight ($s_r = s_\alpha = s_\beta = 1$).

KGE has the same range as NSE, with a value of KGE = 1 indicating optimal model performance. Although they hold the same range, KGE and NSE assess model performance differently and therefore should not be interpreted in the same manner. Each index emphasizes distinct aspects of the model's accuracy and may lead to different conclusions about performance. For example, in the case of NSE, a value of 0.0 or less indicates that it is more accurate to use the observational mean as a predictor instead of simulated values, compared to a benchmark of -0.41 for KGE (Althoff & Rodrigues, 2021).

d. Regression analysis:

R^2 is a statistical metric that represents how well the model can explain the variability in observed data. It describes the degree of collinearity between simulated and observed values and measures the deviation of the observed-simulated regression line from the 1:1 line which indicates an ideal fit between observed and simulated values (Althoff & Rodrigues, 2021).

The R^2 coefficient is calculated as follows:

$$R^2 = \frac{\left[\sum_{i=1}^{n}(O_i - \overline{O}) \cdot (P_i - \overline{P})\right]^2}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2} \cdot \sqrt{\sum_{i=1}^{n}(P_i - \overline{P})^2}}$$

Where $\overline{P}$ is the predicted mean, or the mean of simulated values.

R^2 ranges from 0.0 to 1.0, with values closer to 1 indicating more accurate model accuracy, with $R^2 = 1$ corresponding to optimal model performance and a perfect fit. It is recommended that the regression line gradient (slope) and intercept always be reported alongside R^2, as they provide additional insights to the evaluation of model performance. A good agreement between observed and simulated data corresponds to an intercept closer to 0 and a slope closer to 1 (Moriasi et al., 2015).

e. Percent bias (PBIAS):

PBIAS quantifies the average deviation of simulated values from observed values expressed as a percentage (%), identifying any systematic overestimation or underestimation by the model of an observed variable. PBIAS is calculated as follows:

$$PBIAS = \left[\frac{\sum_{i=1}^{n}(O_i - P_i)}{\sum_{i=1}^{n} O_i}\right] \times 100$$

The optimal value of PBIAS is 0.0, indicating perfect model performance. Lower values indicate accurate model simulation, while high-magnitude values indicate poor model performance. A model bias toward underestimation is indicated by positive PBIAS values, while overestimation bias is explained by negative values (H. V. Gupta et al., 1999).

f. RMSE-observations standard deviation ratio (RSR):

RSR is a model evaluation metric developed alongside the root mean square error (RMSE) to qualify what is considered a low RMSE value based on the standard deviation of the observations. It is calculated as follows:

$$RSR = \frac{RMSE}{STDEV_{obs}} = \frac{\sqrt{\sum_{i=1}^{n}(O_i - P_i)^2}}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2}}$$

RSR ranges from 0.0 to $+\infty$, with RSR = 0.0 indicating zero RMSE and therefore optimal model performance. Lower RSR values indicate lower RMSE and better model accuracy (D. N. Moriasi et al., 2007).

3.3.3. Rating model performance

Model performance was rated according to the following table (**Table 5**). Thresholds for performance metrics and error indices were specified after a thorough review of the literature (Bihon et al., 2024; D. N. Moriasi et al., 2007; Moriasi et al., 2015; Towler et al., 2023).

Table 5. *Model performance rating criteria.*

Rating	NSE	KGE	R^2	PBIAS	RSR
Very good	$0.75 < NSE \leq 1.00$	$0.85 < KGE \leq 1.00$	$0.85 < R^2 \leq 1.00$	$PBIAS < \pm 10$	$0.00 \leq RSR \leq 0.50$
Good	$0.60 < NSE \leq 0.75$	$0.75 < KGE \leq 0.85$	$0.75 < R^2 \leq 0.85$	$\pm 10 \leq PBIAS < \pm 15$	$0.50 < RSR \leq 0.60$
Satisfactory	$0.50 < NSE \leq 0.60$	$0.50 < KGE \leq 0.75$	$0.60 < R^2 \leq 0.75$	$\pm 15 \leq PBIAS < \pm 25$	$0.60 < RSR \leq 0.70$
Unsatisfactory	$NSE \leq 0.50$	$KGE \leq 0.50$	$R^2 \leq 0.60$	$PBIAS \geq \pm 25$	$RSR > 0.70$
Unacceptable	**$NSE < 0.0$**	**$KGE < -0.41$**	**$R^2 < 0.18$**	**$PBIAS \geq \pm 30\%$**	-

4. Triple Collocation-Based Evapotranspiration

Triple collocation (TC) analysis aims to estimate the error variances and correlation of mutually independent products of a geophysical variable in relation to the true value, without the need for a reference dataset. In the case of evapotranspiration (ET), triple collocation methods are used simply to characterize uncertainties from different sources (C. Li et al., 2022), or to fuse ET estimates into one dataset with minimized errors (Park et al., 2023; D. Wang et al., 2024). This methodological framework has been extensively applied in hydrological modeling studies, where TC-based merged ET was integrated in model calibration (X. Guo et al., 2024; C. Wang et al., 2024; Xu et al., 2024)

4.1. ET datasets

Three ET datasets obtained from multiple sources (remote sensing-based, land surface model, reanalysis product) were used in this study:

The ERA5-Land dataset refers to the fifth-generation global climate **reanalysis product** from the European Centre for Medium-Range Weather Forecasts (ECMWF), providing global coverage of land climate variables with a resolution of 0.1° and a temporal coverage from 1950 to present. For this study, monthly averaged ET data (in m/day) was downloaded for the period of 2005-2021[*] (https://cds.climate.copernicus.eu/; accessed March 26, 2025).

The NASA Global Land Data Assimilation System, version 2.1 (GLDAS-2.1) provides global coverage of monthly ET data (in $kg/m^2/s$) averaged from 3-hourly products with a resolution of 0.25° and a temporal coverage from 2000 to present. For this study, ET data was obtained from the Noah **land surface model** (Noah-3.6) from 2005 to 2021 (https://disc.gsfc.nasa.gov/; accessed March 26, 2025).

[*] Incorrect data from September 2022 until February 2024, which is why only data until 2021 were included.

The Global Land Evaporation Amsterdam Model version 4.2a, or GLEAMv4.2a, is a **remote sensing product** that provides global coverage of land evaporation data with a spatial resolution of 0.1° and a temporal coverage of 1980-2023. For this study, monthly ET data (in mm/month) aggregated from daily averages were obtained for the period of 2005-2021 (https://www.gleam.eu/; accessed March 25, 2025).

4.2. Data pre-processing

All datasets were resampled and aligned to a common resolution of 0.1°. All units were converted to get total ET per month in mm. Where applicable, negative values, which refer to condensation, were set as 0, following the suggestions of X. Li et al. (2023).

4.3. Triple collocation

Suppose there are three datasets ET_1, ET_2 and ET_3 (collectively referred to as ET_i), and a true value ET_{true}. In the TC method (McColl et al., 2014), the relationship between each dataset and the true value is assumed to be linear, expressed as:

$$ET_i = \alpha_i + \beta_i ET_{true} + \varepsilon_i$$

Where ε_i is the random error, while β_i and α_i are the slope and intercept of the ordinary least squares. This method requires three independent input datasets and follows three assumptions: (1) zero error cross-correlation (random errors of the three products should be independent; (2) error orthogonality (random errors should be uncorrelated with the truth value); and (3) zero-mean random error (the mathematical expectation of random errors from the three datasets should be zero). Based on these assumptions, the covariance C_{ij} between two different datasets ET_i and ET_j is calculated as:

$$C_{ij} = Cov(ET_i, ET_j) = \begin{cases} \beta_i \beta_j \sigma_{true}^2 \ (i \neq j) \\ {\beta_i}^2 \sigma_{true}^2 + {\sigma_{\varepsilon_i}}^2 \ (i = j) \end{cases}$$

Where σ_{true}^2 is the variance of the truth value and ${\sigma_{\varepsilon_i}}^2$ refers to the variance of each dataset against truth value. Thus, the covariances of each product can be expressed as:

$$\begin{cases} C_{11} = {\beta_1}^2 \sigma_{true}^2 + {\sigma_{\varepsilon_1}}^2 \\ C_{22} = {\beta_2}^2 \sigma_{true}^2 + {\sigma_{\varepsilon_2}}^2 \\ C_{33} = {\beta_3}^2 \sigma_{true}^2 + {\sigma_{\varepsilon_3}}^2 \end{cases}, \begin{cases} C_{12} = \beta_1 \beta_2 \sigma_{true}^2 \\ C_{13} = \beta_1 \beta_3 \sigma_{true}^2 \\ C_{23} = \beta_2 \beta_3 \sigma_{true}^2 \end{cases}$$

And therefore, the error variances of each dataset against the truth value are calculated as:

$$\begin{cases} \sigma_{\varepsilon_1}^2 = C_{11} - \dfrac{C_{12}C_{13}}{C_{23}} \\ \sigma_{\varepsilon_2}^2 = C_{22} - \dfrac{C_{12}C_{23}}{C_{13}} \\ \sigma_{\varepsilon_3}^2 = C_{33} - \dfrac{C_{13}C_{23}}{C_{12}} \end{cases}$$

This research adopts the dual time-space merging approach introduced by Zhou et al. (2021), which takes into consideration spatial and temporal non-stationary errors instead of assuming that random error in the products is statistically stationary. Therefore, both spatial ($\sigma_{\varepsilon_{i,t}}^2$) and temporal ($\sigma_{\varepsilon_{i,s}}^2$) error variances are computed for each dataset, following the methods described above. Then, the spatio-temporal error variance $\sigma_{\varepsilon_{i,t,s}}^2$ is deduced as the weighted average as per the equation below, where N_t and N_s refer to the spatial (per time step) and temporal (per grid cell) sample size, respectively:

$$\sigma_{\varepsilon_{i,t,s}}^2 = \frac{N_t\sigma_{\varepsilon_{i,t}}^2 + N_s\sigma_{\varepsilon_{i,s}}^2}{N_t + N_s}$$

The Least Squares-Based Data Fusion method was adopted to merge the triplet datasets and obtain a fused ET product (ET_m) with minimized errors.

$$ET_m = w_1ET_1 + w_2ET_2 + w_3ET_3$$

Where w_1, w_2, and w_3 are the weights of each of the three collocated datasets, calculated based on the spatio-temporal error variances obtained from TC analysis:

$$w_1 = \frac{\sigma_{\varepsilon_{2,t,s}}^2\sigma_{\varepsilon_{3,t,s}}^2}{\sigma_{\varepsilon_{1,t,s}}^2\sigma_{\varepsilon_{2,t,s}}^2 + \sigma_{\varepsilon_{1,t,s}}^2\sigma_{\varepsilon_{3,t,s}}^2 + \sigma_{\varepsilon_{2,t,s}}^2\sigma_{\varepsilon_{3,t,s}}^2}$$

$$w_2 = \frac{\sigma_{\varepsilon_{1,t,s}}^2\sigma_{\varepsilon_{3,t,s}}^2}{\sigma_{\varepsilon_{1,t,s}}^2\sigma_{\varepsilon_{2,t,s}}^2 + \sigma_{\varepsilon_{1,t,s}}^2\sigma_{\varepsilon_{3,t,s}}^2 + \sigma_{\varepsilon_{2,t,s}}^2\sigma_{\varepsilon_{3,t,s}}^2}$$

$$w_3 = \frac{\sigma_{\varepsilon_{1,t,s}}^2\sigma_{\varepsilon_{2,t,s}}^2}{\sigma_{\varepsilon_{1,t,s}}^2\sigma_{\varepsilon_{2,t,s}}^2 + \sigma_{\varepsilon_{1,t,s}}^2\sigma_{\varepsilon_{3,t,s}}^2 + \sigma_{\varepsilon_{2,t,s}}^2\sigma_{\varepsilon_{3,t,s}}^2}$$

The sum of the weights should be equal to 1.0 to obtain an unbiased ET_m.

4.4. Evaluation of ET_m

The MODIS evapotranspiration product (MOD16A2, Collection 6.1; Running et al., 2021) was used as a reference to evaluate the merged ET product and compare it with the parent datasets. MOD16A2 provides 8-day composite evapotranspiration (ET) data at a spatial resolution of 500 m, with global coverage and temporal availability from 2000 to present. For this study, 8-day ET data from 2010 to 2021 were obtained from the NASA Land

Processes Distributed Active Archive Center (LP DAAC) (https://lpdaac.usgs.gov/; accessed April 24, 2025). The 8-day ET values (kg/m²/8-day) were aggregated to obtain monthly totals. The following metrics were computed to evaluate ET_m: NSE, KGE, R^2, PBIAS, RSR, and Relative Error (RE). A Wilcoxon signed-rank test was performed to assess whether the differences between ET_m and the parent datasets were statistically significant.

5. Results and analysis

5.1. Merged ET dataset

The triple collocation–based merged evapotranspiration dataset (ET_m) generally exhibits improved performance relative to the individual parent datasets, particularly when compared to the reanalysis (ERA5) and land surface model-based (GLDAS) products. Across all evaluation metrics (**Table 6**), ET_m shows higher mean NSE, KGE, and R^2 values, alongside lower RSR and relative error, indicating an overall reduction in random error and improved agreement with the reference dataset. Although NSE values remain low for all products, highlighting the inherent challenges in ET validation, ET_m consistently reduces performance dispersion, as reflected by lower standard deviations across most metrics. Negative NSE values do not necessarily indicate poor ET realism, as time series-based evaluation is highly sensitive to bias and temporal variance, particularly when reference ET exhibits low interannual variability; therefore, relative improvements across datasets and reduced dispersion provide more informative indicators of ET_m performance than absolute NSE magnitudes. The spatial performance of the ET parent and merged datasets is visualized in **Annex A**.

Table 6. *Summary statistics (mean ± SD) of evaluation metrics for merged and parent ET datasets.*

Metric	ET_m	ET_1	ET_2	ET_3
NSE	-2.29 ± 2.98	-3.73 ± 4.36	-4.74 ± 7.00	-1.40 ± 13.99
KGE	-0.03 ± 0.55	-0.24 ± 0.69	-0.32 ± 0.79	0.15 ± 0.46
R^2	0.69 ± 0.22	0.69 ± 0.20	0.66 ± 0.23	0.64 ± 0.23
|PBIAS|	39.23 ± 23.59	50.17 ± 26.48	46.48 ± 33.52	14.94 ± 37.35
RSR	1.60 ± 0.84	1.92 ± 1.00	2.03 ± 1.27	1.33 ± 0.79
RelError	52.20 ± 30.08	61.72 ± 33.76	61.30 ± 36.78	42.43 ± 39.53

Note: *ET_1 = ERA5-Land (reanalysis); ET_2 = GLDAS-Noah v3.6 (land surface model); ET_3 = GLEAM v4.2a (remote sensing–based); ET_m = triple collocation–based merged evapotranspiration dataset.*

Pairwise comparisons further confirm the added value of the merging approach (**Table 7**). ET_m outperforms ERA5 and GLDAS in the majority of evaluated cases across both performance and error metrics, demonstrating the effectiveness of triple collocation in mitigating product-specific biases and uncertainties. In contrast, improvements relative to the remote sensing-based product (GLEAM) are less systematic, suggesting that this dataset already provides comparatively stable ET estimates in the study region. Nevertheless, ET_m

maintains comparable performance to GLEAM while benefiting from reduced bias and improved error balance across datasets.

Table 7. *Percentage of cases where ET_m outperforms individual datasets across evaluation metrics.*

Metric	$\%ET_m > ET_1$	$\%ET_m > ET_2$	$\%ET_m > ET_3$
NSE	90.07	84.45	20.83
KGE	87.23	83.46	21.84
R^2	54.72	68.31	64.69
|PBIAS|	85.23	63.56	15.25
RSR	90.09	84.47	20.83
RelError	88.73	73.79	20.83

Non-parametric Wilcoxon signed-rank tests further confirm these pairwise comparisons, indicating that the merged evapotranspiration product (ET_m) performs significantly better than ERA5 and GLDAS across all evaluated performance and error metrics ($p < 0.05$). In contrast, relative to the remote-sensing-based GLEAM product, ET_m performs significantly better only for R^2, while exhibiting significantly poorer performance for NSE, KGE, PBIAS, RSR, and Relative Error ($p < 0.05$).

Analysis based on performance thresholds supports these findings (**Table 8**). ET_m consistently increases the proportion of cases meeting positive NSE values and higher NSE, KGE, and R^2 thresholds compared to ERA5 and GLDAS, while achieving performance levels similar to GLEAM at higher thresholds. This indicates that the merged product not only improves average performance but also enhances the robustness and consistency of ET estimates.

Table 8. *Percentage of cases meeting performance thresholds across merged and parent ET datasets.*

Metric	Threshold	$\%ET_m$	$\%ET_1$	$\%ET_2$	$\%ET_3$
NSE	**<0**	74.80	80.11	76.29	68.89
	0.000	25.20	19.89	23.71	31.11
	0.500	17.61	13.77	16.70	19.64
	0.800	12.34	7.39	6.84	12.53
	0.900	6.16	2.00	1.52	6.37
KGE	**<-0.4**	52.25	60.48	62.98	35.97
	0.5	19.59	16.64	17.81	25.97
	0.800	8.84	6.48	8.79	7.56
	0.900	3.23	1.00	1.99	2.55
R^2	**<0.5**	0.00	0.00	0.00	0.00
	0.500	79.46	81.08	78.53	74.37
	0.800	38.48	35.37	35.19	25.44
	0.900	18.60	10.98	7.17	17.24

Overall, the results demonstrate that the triple collocation–based merging strategy produces a more balanced and reliable ET dataset by combining complementary strengths of the input products. The resulting ET_m dataset is therefore well suited for subsequent hydrological model calibration and water balance analysis, particularly in regions where uncertainty in individual ET products may propagate into model simulations.

However, it is worth noting that a known limitation of the triple collocation (TC) framework is its sensitivity to violations of underlying assumptions, particularly error independence and negligible cross-correlation among datasets. When these assumptions are not satisfied, or when correlations are weak and sample sizes are limited, TC may yield non-physical estimates such as negative error variances or covariances, which indicate that error magnitudes cannot be reliably resolved rather than representing meaningful uncertainties, e.g. González-Gambau et al. (2020); Su et al. (2014). Previous studies have shown that even modest error cross-correlations or unequal error magnitudes can bias TC solutions and lead to negative variance estimates, especially under finite sampling conditions (González-Gambau et al., 2020; Sjoberg et al., 2021). In this study, negative variance values prevented the direct computation of TC-based weights in some cases. Rather than discarding these cases, equal weighting was applied as a pragmatic and transparent alternative, following similar approaches in previous studies that applied fallback or substituted weighting strategies when TC solutions were non-physical (Wei et al., 2023). While acknowledging that the resulting merged estimates are indicative rather than optimal, valid TC solutions obtained for most cases (≈92.8% for GLEAM, 96.6% for GLDAS, and 85.8% for ERA5) suggest that the merged ET product remains robust overall, although merged values derived under non-physical TC conditions should be interpreted with increased caution.

5.2. Initial SWAT+ performance

Initial model performance in simulating streamflow is acceptable at all validation sites (**Table 9**). Model performance is rated as satisfactory according to all criteria, except for NSE, PBIAS, and RSR which yield good to very good performance at the upstream sites. The negative (S1) and positive (S2 and S3) values of PBIAS indicate model overprediction and underprediction, respectively.

Table 9. *Performance metrics for initial streamflow simulation prior to calibration.*

	Downstream site	**Upstream sites**	
	Teo (S1)	Deza (S2)	Furelos (S3)
NSE	0.54	0.62	0.64
KGE	0.68	0.66	0.64
R2	0.69	0.62	0.65
PBIAS	-23.07	+5.68	+12.88
RSR	0.67	0.62	0.60

The SWAT+ model overestimates mean flow for S1 (μ_{obs}= 58.2 m^3/s < μ_{sim}= 72.4 m^3/s) but predicts the mean flow for S2 (μ_{obs}= 15.4 m^3/s ≈ μ_{sim}= 14.5 m^3/s) and S3 (μ_{obs}= 5.6 m^3/s ≈ μ_{sim}= 5 m^3/s) relatively well. The same trend is observed for the median. Furthermore, the model significantly underestimates minimum and maximum flow for all sites, while the standard deviation is overestimated at S1 (SD_{obs}= 62.16 m^3/s < SD_{sim}= 70.36 m^3/s) and underestimated at both S2 (SD_{obs}= 20.70 m^3/s < SD_{sim}= 15.38 m^3/s) and S3 (SD_{obs}= 7.41 m^3/s < SD_{sim}= 5.48 m^3/s).

The time series plots show that SWAT+ is able to simulate the trends in streamflow relatively well, with significant underestimation of peak flows in all sites. This is also apparent in the scatter plots and flow duration curve (FDC). FDC plots reveals that this underprediction was more extreme for the upstream sites. As for normal flow conditions, the model simulates streamflow considerably well at sites S2 and S3. In contrast, for S1, the model tends to overestimate normal flow. All plots can be found in **Annex B**.

For evapotranspiration, model performance varies across the landscape units (LSUs). Overall, the model significantly underpredicts ET over the whole watershed (**Table 10**). NSE scores range from -0.45 to 0.76 with a mean of 0.20, indicating unsatisfactory model performance. The same is observed for PBIAS (-48.2 to -15.2, $PBIAS_{mean}$ = -32.82) and RSR (0.49 to 1.19, RSR_{mean} = 0.87). However, for KGE (0.32 to 0.80, KGE = 0.56) and R^2 (0.31 to 0.87, R^2_{mean} = 0.61), results are satisfactory.

Table 10. *Summary statistics for observed and simulated ET, averaged across all LSUs.*

	Observed	**Simulated**
Mean	62.98	41.21
Min	14.77	4.13
Max	122.43	100.42
Median	61.89	33.16
SD	34.17	27.79

Overall, the uncalibrated SWAT+ model demonstrates an acceptable baseline performance for both streamflow and evapotranspiration, indicating that the model structure, input data, and watershed representation are internally consistent and free of major setup deficiencies. Based on summary statistics, performance metrics, and graphical diagnostics, the model successfully captures the overall magnitude, temporal patterns, and dominant dynamics of streamflow across all stations, despite evident biases in peak flows and low-flow extremes. Similarly, while evapotranspiration is systematically underestimated across landscape units, correlation-based metrics confirm that the spatial and temporal variability of ET is reasonably represented. These results indicate that the observed discrepancies reflect correctable parameter-related biases rather than structural or conceptual errors in the model configuration

and therefore provide a sound and defensible basis for proceeding with targeted single-variable and multivariate calibration strategies in the following section.

5.3. SWAT+ calibration

Sensitive parameters ($p < 0.05$) for each calibration scheme are listed in **Table 11**. Across all calibration schemes, sensitivity patterns consistently highlight the dominant role of soil–vegetation interaction parameters in controlling both streamflow and evapotranspiration. Parameters related to soil water storage and redistribution (awc.sol, esco.hru) and canopy processes (canmx.hru) appear among the most sensitive in all schemes, indicating that vertical water fluxes exert first-order control on basin hydrological behavior. In contrast, routing and groundwater parameters exhibit higher sensitivity under streamflow-based calibration, reflecting their stronger influence on flow timing and magnitude, particularly at upstream stations. This consistency across calibration strategies supports the internal coherence of the model structure and suggests that differences in performance arise from how these shared controls are weighted rather than from fundamentally different parameter sensitivities.

Table 11. *Sensitive parameters identified for each calibration scheme, ranked in order of sensitivity.*

	SC-Q	SC-ET	MC-QET
Sensitive parameters	chn.rte awc.sol cn3_swf.hru perco.hru cn2.hru latq_co.hru surlag.bsn flo_min.aqu esco.hru canmx.hru	awc.sol canmx.hru latq_co.hru esco.hru cn2.hru	awc.sol canmx.hru esco.hru latq_co.hru snofall_tmp.hru cn2.hru evlai.bsn cn3_swf.hru perco.hru

To ensure consistency, NSE is selected as the objective function for all calibration schemes, with a behavioral threshold of 0.60. Validation results of the calibrated model can be found in **Table 12** (streamflow) and **Table 13** (ET). Plots are included in **Annex C**.

The multivariate calibration approach clearly outperforms the single-variate schemes only at the downstream site, although all the methods yield good to very good results. For ET specifically, SC-ET and MC-QET yield the same results, and both outperform SC-Q. These results show that while single-variate calibration based on streamflow can optimize the targeted variable, integration of ET proves useful to get more reliable water balance predictions. Furthermore, ET-only calibration also proves reliable, which is specifically important for watersheds with limited or no gauging data.

Importantly, these findings demonstrate that differences among calibration schemes primarily affect spatial transferability and process representation rather than overall model stability, reinforcing the suitability of multivariate calibration when the objective extends beyond site-specific streamflow reproduction.

Table 12. *Performance metrics of streamflow simulation under each calibration scheme across all validation sites.*

			NSE	KGE	R^2	PBIAS	RSR
Downstream site	**S1**	**SC-Q**	0.79	0.71	0.89	-22.21	0.46
		SC-ET	0.70	0.72	0.82	-5.16	0.55
		MC-QET	0.82	0.81	0.87	-4.51	0.43
Upstream sites	**S2**	**SC-Q**	0.86	0.83	0.87	6.35	0.37
		SC-ET	0.75	0.73	0.78	22.35	0.50
		MC-QET	0.79	0.70	0.83	23.27	0.45
	S3	**SC-Q**	0.82	0.84	0.83	12.72	0.43
		SC-ET	0.71	0.66	0.75	26.01	0.54
		MC-QET	0.79	0.70	0.83	26.55	0.46

Table 13. *Performance metrics of ET simulation under each calibration scheme, averaged across all LSUs.*

	NSE_{mean}	KGE_{mean}	R^2_{mean}	$PBIAS_{mean}$*	RSR_{mean}
SC-Q	0.66	0.70	0.88	22.97	0.57
SC-ET	0.89	0.89	0.92	6.56	0.32
MC-QET	0.90	0.89	0.93	6.24	0.31

**computed as the mean of absolute PBIAS values*

The spatial pattern of model performance reveals distinct scale-dependent responses to the calibration strategies. At the downstream station, the multivariate calibration (MC-QET) yields superior performance compared to single-variable approaches, indicating that the inclusion of evapotranspiration effectively constrains long-term water balance and reduces compensatory errors between runoff generation and groundwater processes at larger spatial scales. Conversely, at upstream stations, where hydrological response is more strongly controlled by short-term rainfall–runoff dynamics and reduced storage integration, streamflow-only calibration (SC-Q) remains the most effective strategy. These results suggest that the added value of multivariate calibration increases with basin scale and hydrological complexity, while single-variable approaches may remain adequate in smaller, faster-responding sub-catchments.

5.4. Uncertainty analysis

Uncertainty metrics reveal a clear trade-off between prediction coverage and uncertainty band thickness across calibration strategies. Streamflow simulations under SC-Q achieve comparatively higher p-factor values, indicating better coverage of observations within the 95PPU envelope, although at the expense of slightly wider uncertainty bands (**Table 14**). In

contrast, MC-QET yields consistently lower p-factors but also markedly lower r-factors across all stations, reflecting narrower prediction intervals and reduced parameter equifinality. This behavior indicates that the inclusion of evapotranspiration in calibration constrains the feasible parameter space, limiting compensatory effects between hydrological fluxes while improving the precision of simulated responses.

Table 14. *Uncertainty analysis results for streamflow simulation under each calibration scheme across all validation sites.*

			Calibration		Validation	
			p-factor	**r-factor**	**p-factor**	**r-factor**
Downstream site	**S1**	**SC-Q**	0.74	0.84	0.74	0.96
		MC-QET	0.58	0.58	0.49	0.54
Upstream sites	**S2**	**SC-Q**	0.55	0.49	0.71	0.58
		MC-QET	0.38	0.35	0.5	0.43
	S3	**SC-Q**	0.62	0.48	0.59	0.52
		MC-QET	0.4	0.39	0.46	0.37

Spatial differences in uncertainty behavior further highlight scale effects within the basin. At the downstream station, where flow integrates responses from multiple upstream processes, uncertainty envelopes remain relatively stable across calibration schemes, suggesting robust constraint of dominant hydrological controls. Upstream stations exhibit lower p-factors overall, particularly under multivariate calibration, reflecting higher sensitivity to localized runoff generation and reduced averaging of errors. These results indicate that multivariate calibration promotes greater uncertainty reduction at larger spatial scales, while smaller sub-catchments retain inherently higher uncertainty associated with rapid hydrological response and limited storage buffering.

For evapotranspiration, relatively low p-factor values were obtained under both SC-ET and MC-QET, despite strong performance metrics (**Table 15**). This outcome reflects the combined effects of spatial averaging across landscape units, monthly aggregation, and remaining uncertainties in remotely sensed ET estimates used as calibration targets. The consistently low r-factors, however, indicate that uncertainty bands remain narrow and stable, suggesting precise ET simulations even when full observational coverage cannot be achieved. These patterns support the interpretation that ET uncertainty is dominated by observational and structural limitations rather than by excessive model parameter variability.

Table 15. *Uncertainty analysis results for ET simulation under each calibration scheme, averaged across all LSUs.*

	Calibration		Validation	
	p-factor$_{mean}$	**r-factor$_{mean}$**	**p-factor$_{mean}$**	**r-factor$_{mean}$**
SC- ET	0.31	0.4	0.27	0.39
MC-QET	0.31	0.39	0.28	0.4

Overall, the uncertainty analysis demonstrates that the SWAT+ simulations are characterized by well-defined and interpretable uncertainty structures across calibration strategies. While full observational coverage is not always achieved, particularly for evapotranspiration and at upstream stations, the generally low r-factors across variables and schemes indicate constrained prediction intervals and limited parameter equifinality. This suggests that uncertainty arises primarily from data limitations and inherent process variability rather than from model instability or over-parameterization. Taken together with the strong performance metrics previously reported, the uncertainty results support the reliability and physical plausibility of the calibrated model and confirm that the derived comparisons between single-variable and multivariate calibration strategies are robust and meaningful.

6. References

Abbas, S. A., Bailey, R. T., White, J. T., Arnold, J. G., White, M. J., ÄaŒerkasova, N., & Gao, J. (2024). A framework for parameter estimation, sensitivity analysis, and uncertainty analysis for holistic hydrologic modeling using SWAT+. *Hydrology and Earth System Sciences*, *28*(1), 21–48. https://doi.org/10.5194/HESS-28-21-2024

Abbaspour, K. C. (2015). *SWAT-CUP: SWAT Calibration and Uncertainty Programs - A User Manual.*

Abbaspour, K. C., Johnson, C. A., & van Genuchten, M. Th. (2004). Estimating Uncertain Flow and Transport Parameters Using a Sequential Uncertainty Fitting Procedure. *Vadose Zone Journal*, *3*(4), 1340–1352. https://doi.org/10.2136/vzj2004.1340

Abbaspour, K. C., Rouholahnejad, E., Vaghefi, S., Srinivasan, R., Yang, H., & Kløve, B. (2015). A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale SWAT model. *Journal of Hydrology*, *524*, 733–752. https://doi.org/10.1016/J.JHYDROL.2015.03.027

Althoff, D., & Rodrigues, L. N. (2021). Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment. *Journal of Hydrology*, *600*, 126674. https://doi.org/10.1016/j.jhydrol.2021.126674

Aragaw, H. M., & Mishra, S. K. (2022). Multi-site multi-objective calibration of SWAT model using a large dataset for improved performance in Ethiopia. *Arabian Journal of Geosciences*, *15*, 320. https://doi.org/10.1007/S12517-022-09602-5

Arnold, J. G., Srinivasan, R., Muttiah, R. S., & Williams, J. R. (1998). LARGE AREA HYDROLOGIC MODELING AND ASSESSMENT PART I: MODEL DEVELOPMENT. *Journal of the American Water Resources Association*, *34*(1), 73–89. https://doi.org/10.1111/J.1752-1688.1998.TB05961.X

Augas de Galicia. (2025, February 19). *Embalses de Galicia-Costa*. https://augasdegalicia.xunta.gal/seccion-tema/c/Control_caudais_reservas?content=rede-encoros/seccion.html&sub=subseccion2/

Bieger, K., Arnold, J. G., Rathjens, H., White, M. J., Bosch, D. D., Allen, P. M., Volk, M., & Srinivasan, R. (2017). Introduction to SWAT+, A Completely Restructured Version of the Soil and Water Assessment Tool. *Journal of the American Water Resources Association*, *53*(1), 115–130. https://doi.org/10.1111/1752-1688.12482

Bihon, Y. T., Lohani, T. K., Ayalew, A. T., Neka, B. G., Mohammed, A. K., Geremew, G. B., & Ayele, E. G. (2024). Performance evaluation of various hydrological models with respect to hydrological responses under climate change scenario: a review. *Cogent Engineering*, *11*(1). https://doi.org/10.1080/23311916.2024.2360007

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., & Montanari, A. (2012). Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. *Physics and Chemistry of the Earth, Parts A/B/C*, *42–44*, 70–76. https://doi.org/10.1016/j.pce.2011.07.037

Castellanos-Osorio, G., López-Ballesteros, A., Pérez-Sánchez, J., & Senent-Aparicio, J. (2023). Disaggregated monthly SWAT+ model versus daily SWAT+ model for estimating environmental flows in Peninsular Spain. *Journal of Hydrology*, *623*, 129837. https://doi.org/10.1016/J.JHYDROL.2023.129837

CNIG. (2025, February 21). *Modelo Digital del Terreno de 1ª cobertura (2009-2015) con paso de malla de 25 metros (MDT25) de España*. https://centrodedescargas.cnig.es/CentroDescargas/modelos-digitales-elevaciones

D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, & T. L. Veith. (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*, *50*(3), 885–900. https://doi.org/10.13031/2013.23153

FAO, & IIASA. (2023). *Harmonized World Soil Database Version 2.0*. https://doi.org/10.4060/cc3823en

Farinango, G., Álvarez-Vázquez, M. Á., & Prego, R. (2023). Trace Element Patterns in Heterogeneous Land–Sea Sediments: A Comprehensive Study of the Ulla–Arousa System (SW Europe). *Geosciences*, *13*(10), 292. https://doi.org/10.3390/GEOSCIENCES13100292

Franco, A. C. L., & Bonumá, N. B. (2017). Multi-variable SWAT model calibration with remotely sensed evapotranspiration and observed flow. *RBRH*, *22*(0). https://doi.org/10.1590/2318-0331.011716090

González-Gambau, V., Turiel, A., González-Haro, C., Martínez, J., Olmedo, E., Oliva, R., & Martín-Neira, M. (2020). Triple Collocation Analysis for Two Error-Correlated Datasets: Application to L-Band Brightness Temperatures over Land. *Remote Sensing*, *12*(20), 3381. https://doi.org/10.3390/rs12203381

Guarracino, L. (2007). Estimation of saturated hydraulic conductivity Ks from the van Genuchten shape parameter α. *Water Resources Research*, *43*(11). https://doi.org/10.1029/2006WR005766

Guo, J., & Su, X. (2019). Parameter sensitivity analysis of SWAT model for streamflow simulation with multisource precipitation datasets. *Hydrology Research*, *50*(3), 861–877. https://doi.org/10.2166/NH.2019.083

Guo, X., Wu, Z., Fu, G., & He, H. (2024). A multi-variable calibration framework at the grid scale for integrating streamflow with evapotranspiration data to improve the simulation of distributed hydrological model. *Journal of Hydrology: Regional Studies*, *55*, 101944. https://doi.org/10.1016/j.ejrh.2024.101944

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1999). Status of Automatic Calibration for Hydrologic Models: Comparison with Multilevel Expert Calibration. *Journal of Hydrologic Engineering*, *4*(2), 135–143. https://doi.org/10.1061/(ASCE)1084-0699(1999)4:2(135)

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 of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Harmel, R. D., Smith, P. K., Migliaccio, K. W., Chaubey, I., Douglas-Mankin, K. R., Benham, B., Shukla, S., Muñoz-Carpena, R., & Robson, B. J. (2014). Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: A review and recommendations. *Environmental Modelling & Software*, *57*, 40–51. https://doi.org/10.1016/j.envsoft.2014.02.013

Koltsida, E., & Kallioras, A. (2022). Multi-Variable SWAT Model Calibration Using Satellite-Based Evapotranspiration Data and Streamflow. *Hydrology*, *9*(7), 112. https://doi.org/10.3390/hydrology9070112

Li, C., Yang, H., Yang, W., Liu, Z., Jia, Y., Li, S., & Yang, D. (2022). Error characterization of global land evapotranspiration products: Collocation-based approach. *Journal of Hydrology*, *612*, 128102. https://doi.org/10.1016/j.jhydrol.2022.128102

Li, X., Zhang, W., Vermeulen, A., Dong, J., & Duan, Z. (2023). Triple collocation-based merging of multi-source gridded evapotranspiration data in the Nordic Region.

Agricultural and Forest Meteorology, *335*, 109451. https://doi.org/10.1016/j.agrformet.2023.109451

Makumbura, R. K., Gunathilake, M. B., Samarasinghe, J. T., Confesor, R., Muttil, N., & Rathnayake, U. (2022). Comparison of Calibration Approaches of the Soil and Water Assessment Tool (SWAT) Model in a Tropical Watershed. *Hydrology*, *9*(10), 183. https://doi.org/10.3390/HYDROLOGY9100183

McColl, K. A., Vogelzang, J., Konings, A. G., Entekhabi, D., Piles, M., & Stoffelen, A. (2014). Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. *Geophysical Research Letters*, *41*(17), 6229–6236. https://doi.org/10.1002/2014GL061322

MeteoGalicia. (2025, February 24). *Lista de estaciónes*. MeteoGalicia. https://www.meteogalicia.gal/web/observacion/rede

MITECO. (2021). *FOTO FIJA 2021. Escala 1:50.000 y 1:25.000 [Dataset]*. Ministerio para la Transición Ecológica y el Reto Demográfico. https://www.miteco.gob.es/es/biodiversidad/temas/inventarios-nacionales/mapa-forestal-espana/foto_fija_mfe.html

Moriasi, D. N., Gitau, M. W., Pai, N., & Daggupati, P. (2015). Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. *Transactions of the ASABE*, *58*(6), 1763–1785. https://doi.org/10.13031/trans.58.10715

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nazari-Sharabian, M., Taheriyoun, M., & Karakouzian, M. (2020). Sensitivity analysis of the DEM resolution and effective parameters of runoff yield in the SWAT model: a case study. *Journal of Water Supply: Research and Technology-Aqua*, *69*(1), 39–54. https://doi.org/10.2166/aqua.2019.044

Nguyen, T. (2024). *tamnva/R-SWAT: v.4.01 (v4.01)*. Zenodo. https://doi.org/10.5281/zenodo.13625844

Nguyen, T. V., Dietrich, J., Dang, T. D., Tran, D. A., Van Doan, B., Sarrazin, F. J., Abbaspour, K., & Srinivasan, R. (2022). An interactive graphical interface tool for parameter calibration, sensitivity analysis, uncertainty analysis, and visualization for the Soil and Water Assessment Tool. *Environmental Modelling & Software*, *156*, 105497. https://doi.org/10.1016/J.ENVSOFT.2022.105497

Oliveira, A. R., Ramos, T. B., Simionesei, L., Pinto, L., & Neves, R. (2020). Sensitivity Analysis of the MOHID-Land Hydrological Model: A Case Study of the Ulla River Basin. *Water*, *12*(11), 3258. https://doi.org/10.3390/W12113258

Park, J., Baik, J., & Choi, M. (2023). Triple collocation-based multi-source evaporation and transpiration merging. *Agricultural and Forest Meteorology*, *331*, 109353. https://doi.org/10.1016/j.agrformet.2023.109353

Post, D. F., Fimbres, A., Matthias, A. D., Sano, E. E., Accioly, L., Batchily, A. K., & Ferreira, L. G. (2000). Predicting Soil Albedo from Soil Color and Spectral Reflectance Data. *Soil Science Society of America Journal*, *64*(3), 1027–1034. https://doi.org/10.2136/SSSAJ2000.6431027X

Running, S., Mu, Q., Zhao, M., & Moreno, A. (2021). MODIS/Terra Net Evapotranspiration Gap-Filled 8-Day L4 Global 500m SIN Grid V061. In *NASA EOSDIS Land Processes Distributed Active Archive Center*. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD16A2GF.061

Senent-Aparicio, J., Jimeno-Sáez, P., López-Ballesteros, A., Giménez, J. G., Pérez-Sánchez, J., Cecilia, J. M., & Srinivasan, R. (2021). Impacts of SWAT weather generator statistics from high-resolution datasets on monthly streamflow simulation over Peninsular Spain. *Journal of Hydrology: Regional Studies*, *35*, 100826. https://doi.org/10.1016/j.ejrh.2021.100826

Sharpley, A. N., & Williams, J. R. (1990). *EPIC—Erosion/Productivity Impact Calculator: 1. Model Documentation (Technical Bulletin No. 1768)*.

Singh, L., & Saravanan, S. (2022). Assessing streamflow modeling using single and multi-site calibration approach on Bharathpuzha catchment, India: a case study. *Modeling Earth Systems and Environment*, *8*(3), 4135–4148. https://doi.org/10.1007/S40808-022-01353-2/FIGURES/7

Sjoberg, J. P., Anthes, R. A., & Rieckh, T. (2021). The Three-Cornered Hat Method for Estimating Error Variances of Three or More Atmospheric Datasets. Part I: Overview and Evaluation. *Journal of Atmospheric and Oceanic Technology*, *38*(3), 555–572. https://doi.org/10.1175/JTECH-D-19-0217.1

Su, C., Ryu, D., Crow, W. T., & Western, A. W. (2014). Beyond triple collocation: Applications to soil moisture monitoring. *Journal of Geophysical Research: Atmospheres*, *119*(11), 6419–6439. https://doi.org/10.1002/2013JD021043

Towler, E., Foks, S. S., Dugger, A. L., Dickinson, J. E., Essaid, H. I., Gochis, D., Viger, R. J., & Zhang, Y. (2023). Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States.

Hydrology and Earth System Sciences, *27*(9), 1809–1825. https://doi.org/10.5194/hess-27-1809-2023

Wang, C., Mao, H., Nemoto, T., He, Y., Hu, J., Li, R., Wu, Q., Wang, M., Song, X., & Duan, Z. (2024). An Adaptive Process-Wise Fitting Approach for Hydrological Modeling Based on Streamflow and Remote Sensing Evapotranspiration. *Water*, *16*(23), 3446. https://doi.org/10.3390/w16233446

Wang, D., Liu, S., & Wang, D. (2024). Triple Collocation-Based Uncertainty Analysis and Data Fusion of Multi-Source Evapotranspiration Data Across China. *Atmosphere*, *15*(12), 1410. https://doi.org/10.3390/atmos15121410

Wang, L., & Liu, H. (2006). An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. *International Journal of Geographical Information Science*, *20*(2), 193–213. https://doi.org/10.1080/13658810500433453

Wei, L., Jiang, S., Ren, L., Yuan, S., Liu, Y., Yang, X., Wang, M., Zhang, L., Yu, H., & Duan, Z. (2023). An Extended Triple Collocation Method With Maximized Correlation for Near Global-Land Precipitation Fusion. *Geophysical Research Letters*, *50*(24). https://doi.org/10.1029/2023GL105120

Williams, J. R. (1995). Chapter 25. The EPIC Model. In *Computer Models of Watershed Hydrology* (pp. 909–1000). Water Resources Publications.

Xu, Z., Liu, J., Wu, Z., & Guo, X. (2024). Enhancing streamflow simulation accuracy in ungauged catchments via parameter calibration with triple collocation-based merged evapotranspiration and streamflow features. *Journal of Hydrology*, *639*, 131627. https://doi.org/10.1016/j.jhydrol.2024.131627

Zhou, J., Crow, W. T., Wu, Z., Dong, J., He, H., & Feng, H. (2021). A triple collocation-based 2D soil moisture merging methodology considering spatial and temporal non-stationary errors. *Remote Sensing of Environment*, *263*, 112509. https://doi.org/10.1016/j.rse.2021.112509

7. Acknowledgements

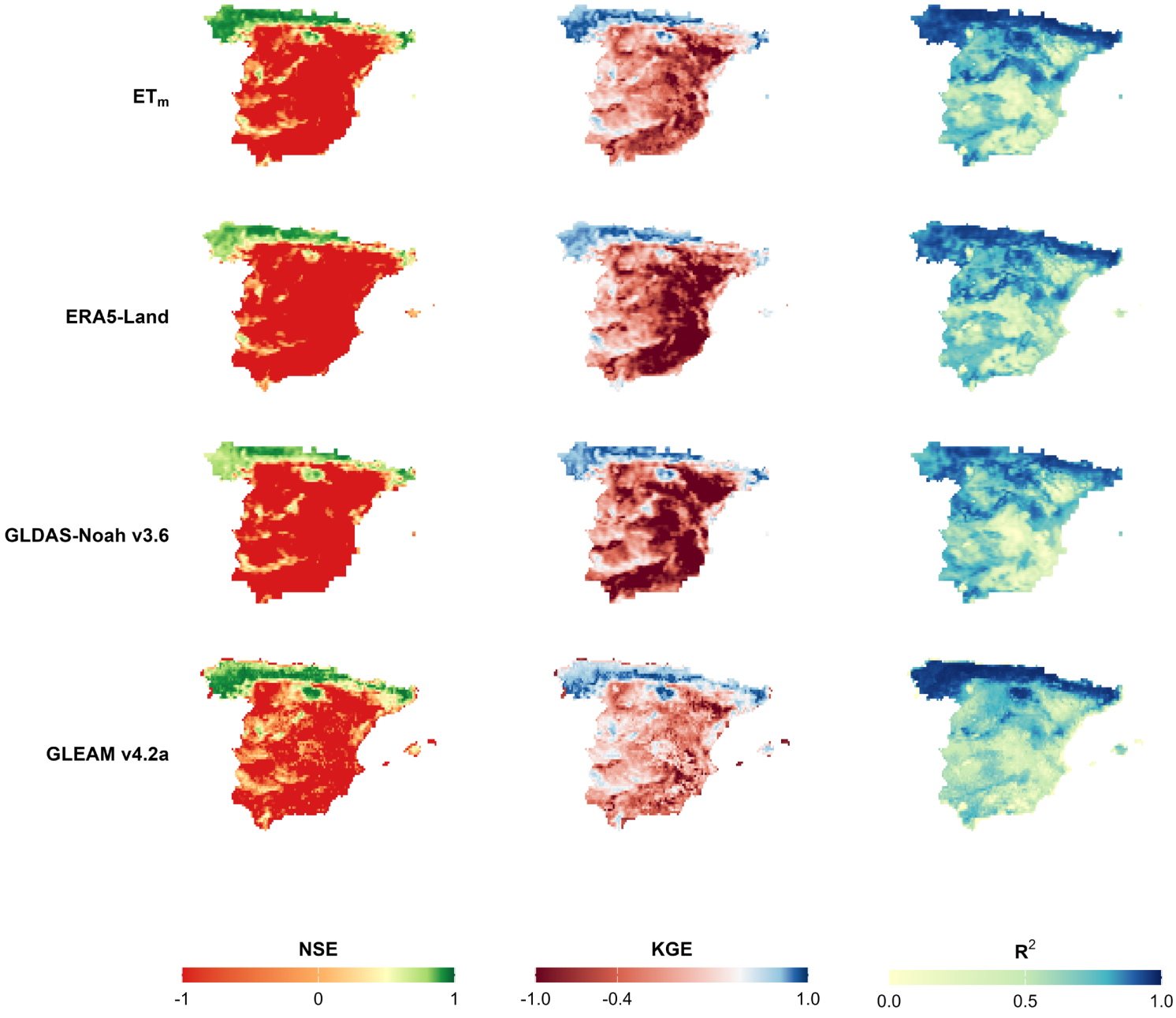
This work was supported by the CHIST-ERA grant CHIST-ERA-23-MultiGIS-06, project PCI2025-163208 funded by MICIU/AEI /10.13039/501100011033 and co-financed by the European Union.

8. Data availability

All data, scripts, and code used and produced in this study are available from the author upon request.

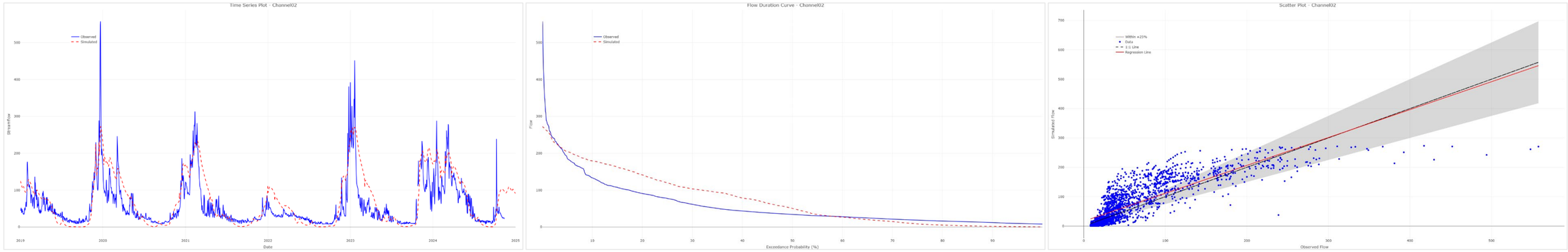
9. Annexes

Annex A: Validation Maps (Parent vs. Merged ET Datasets)

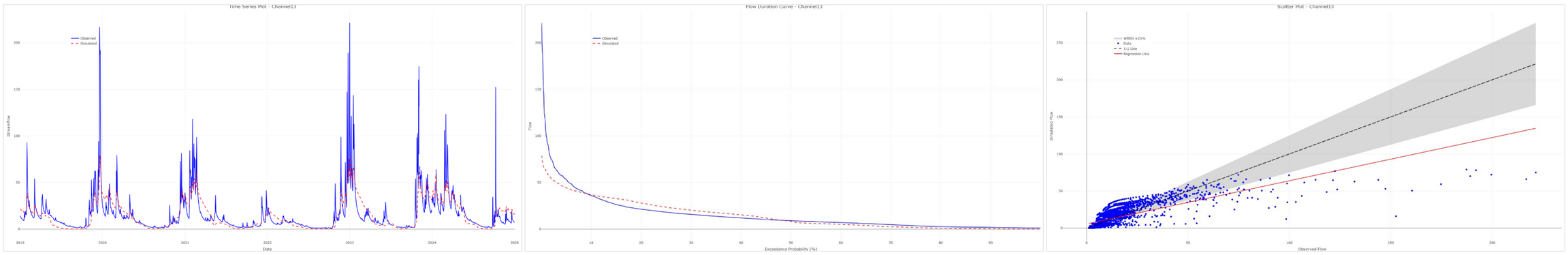


Annex B: Streamflow Validation Plots (Initial SWAT+ Model Performance)

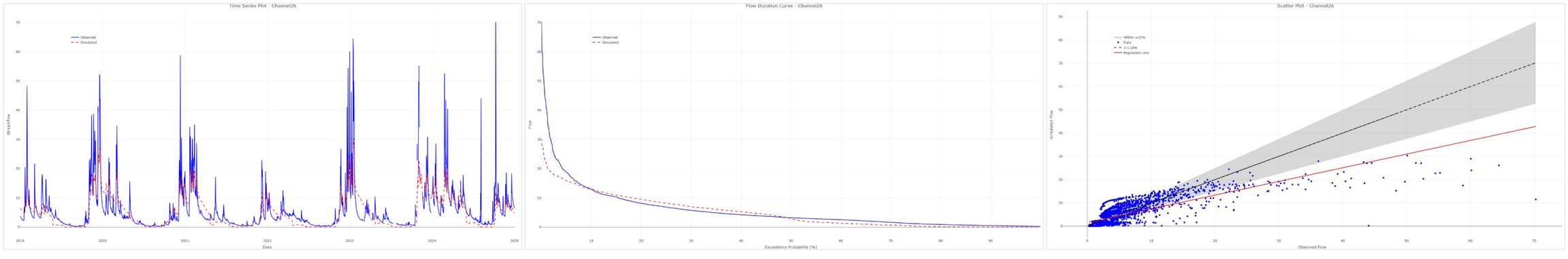
B.1. Channel02: Teo station (downstream)



B.2. Channel13: Deza station (upstream)



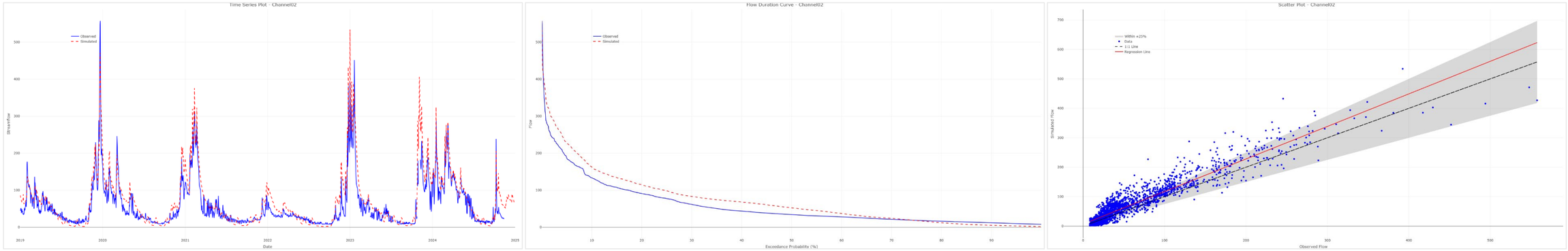
B.3. Channel26: Furelos stations (upstream)



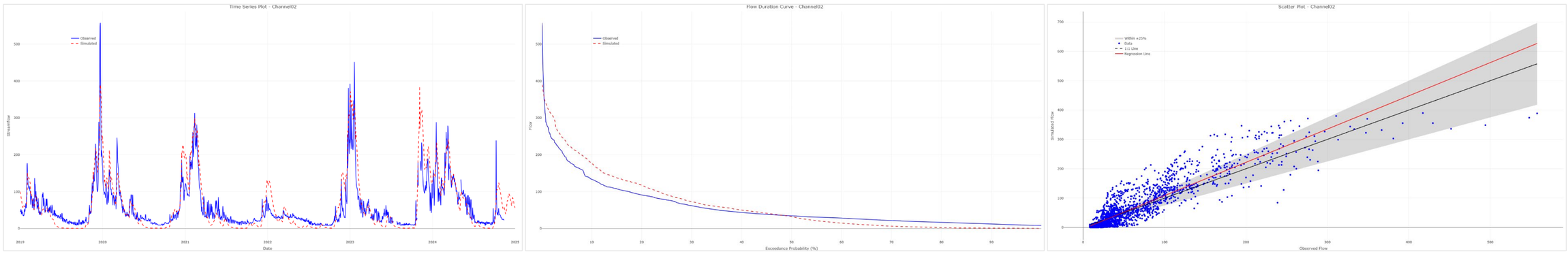
Annex C: Streamflow Validation Plots (Calibrated SWAT + Model Performance)

C.1. Channel02: Teo station (downstream)

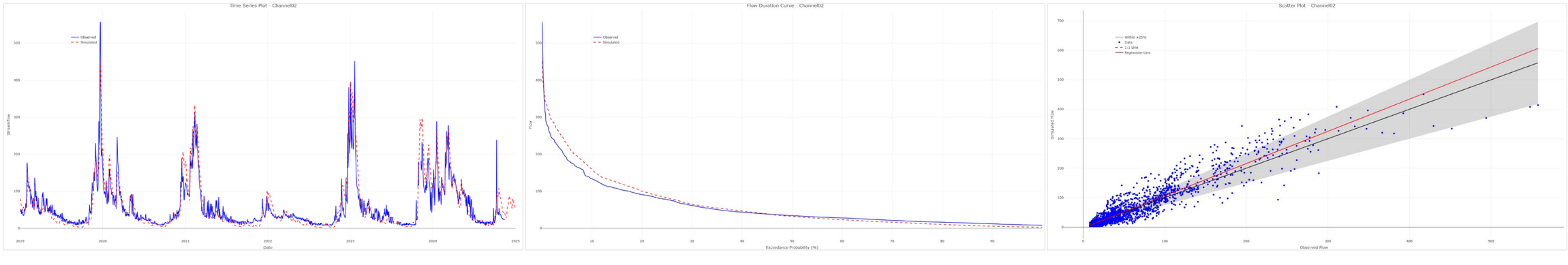
SC–Q Calibration (Q-only)



SC–ET Calibration (ET-only)

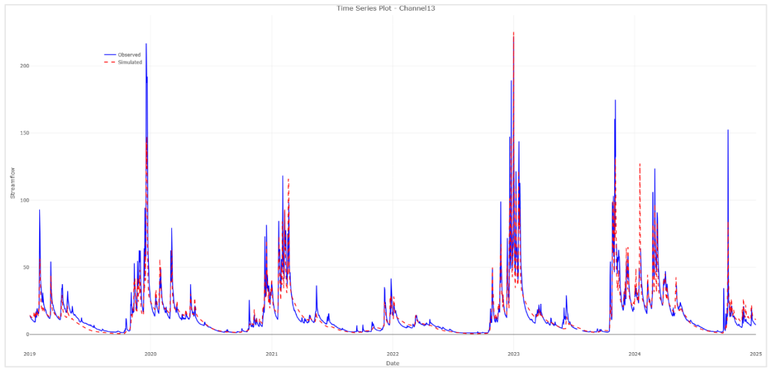


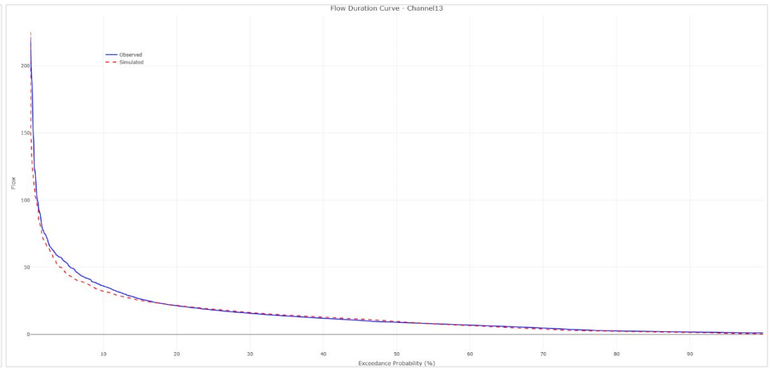
MC–QET Calibration (Q–ET)

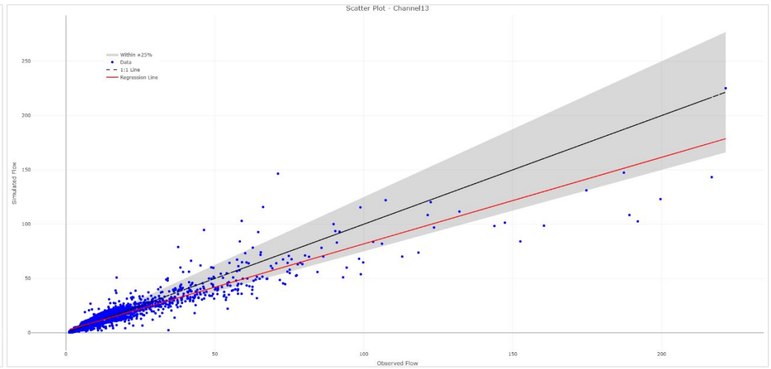


C.2. Channel13: Deza station (upstream)

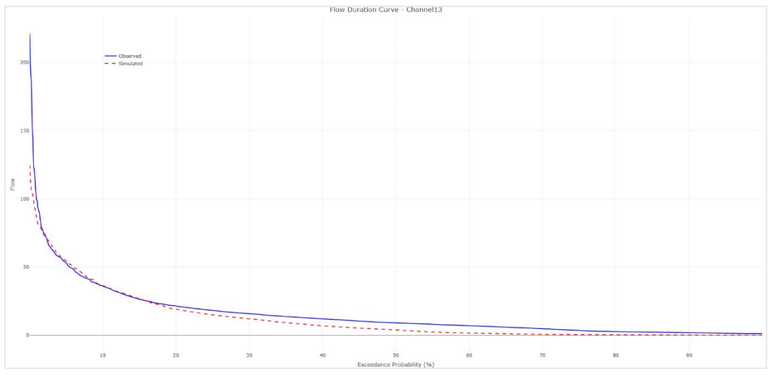
SC–Q Calibration (Q-only)

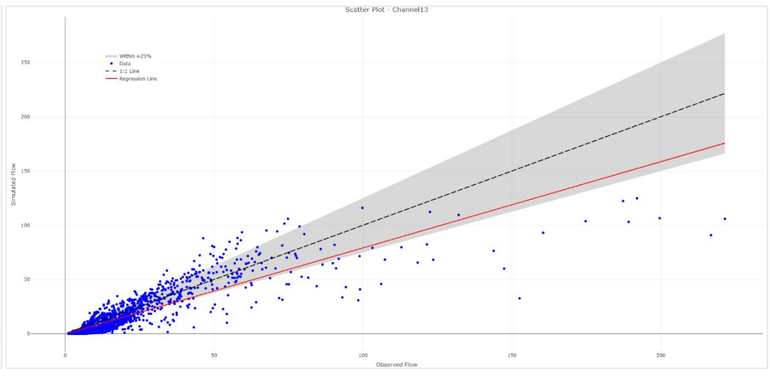


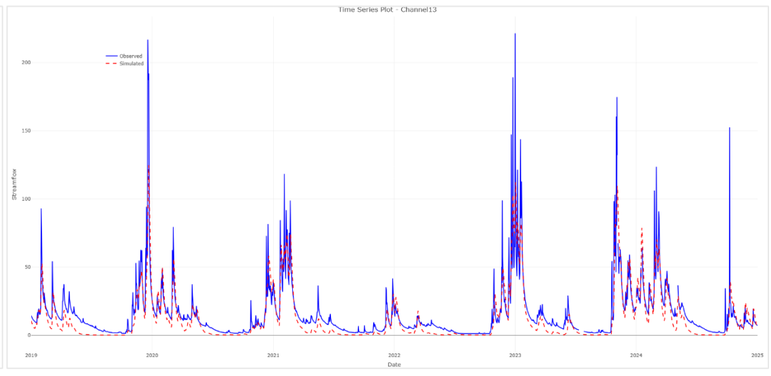




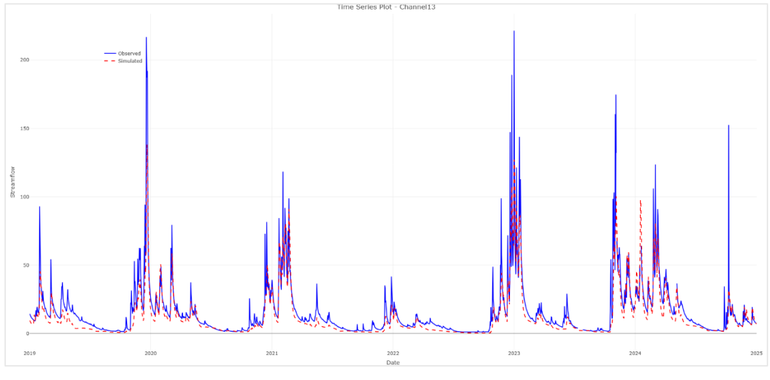
SC–ET Calibration (ET-only)

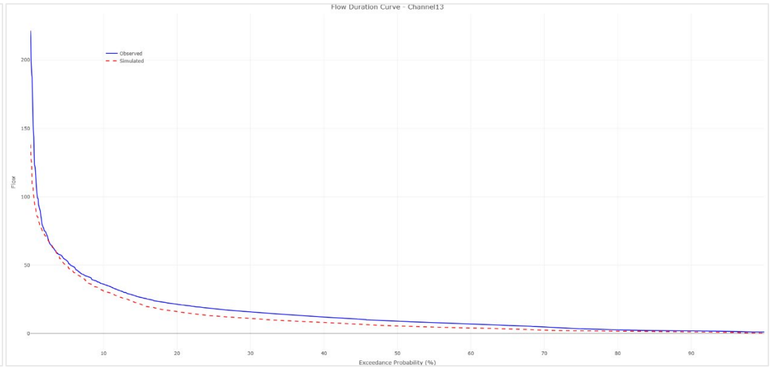


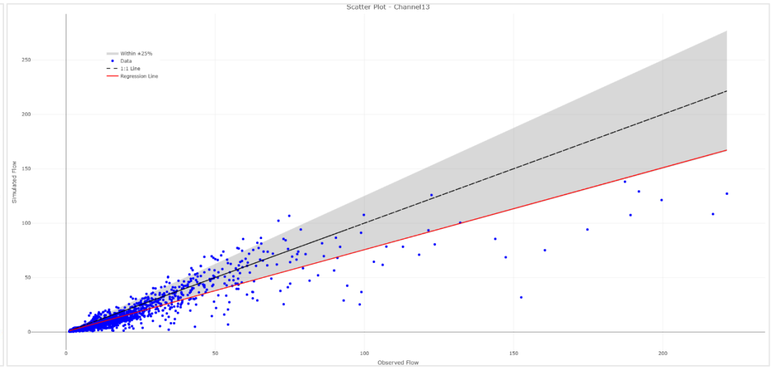




MC–QET Calibration (Q–ET)

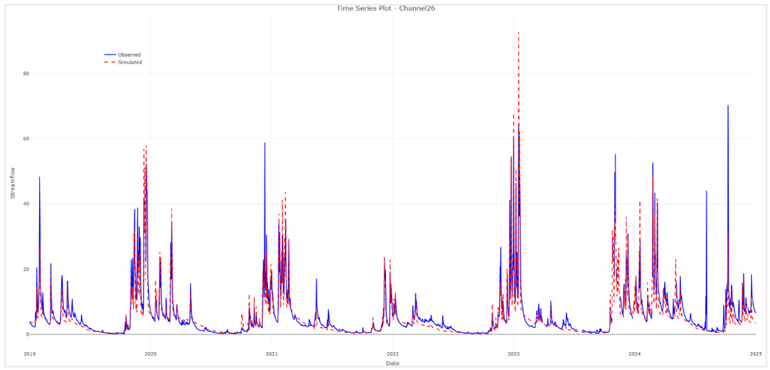


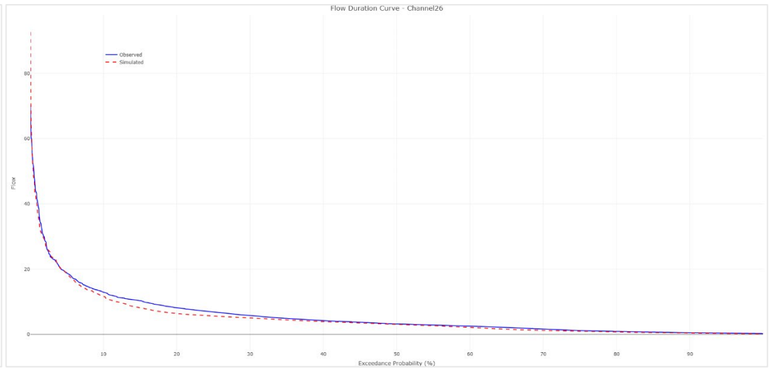


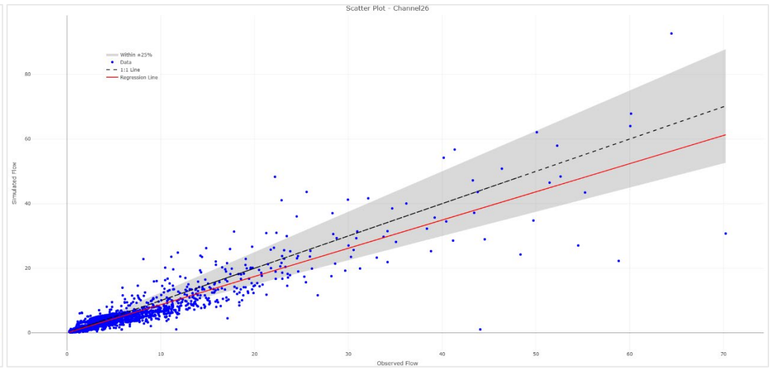


C.3. Channel26: Furelos stations (upstream)

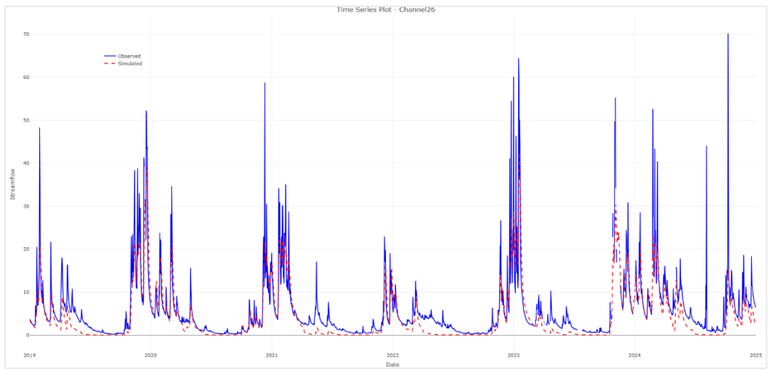
SC–Q Calibration (Q-only)

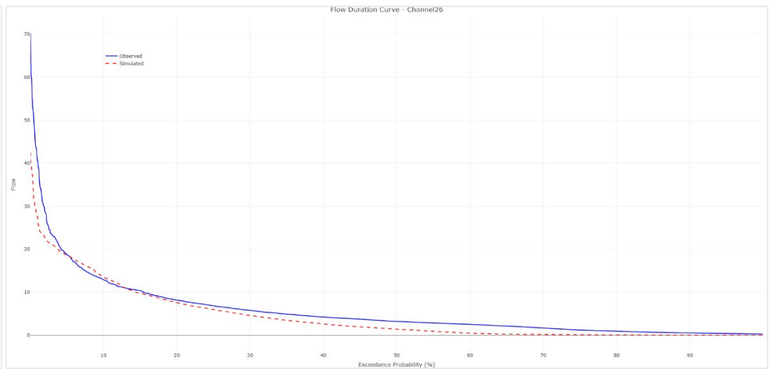


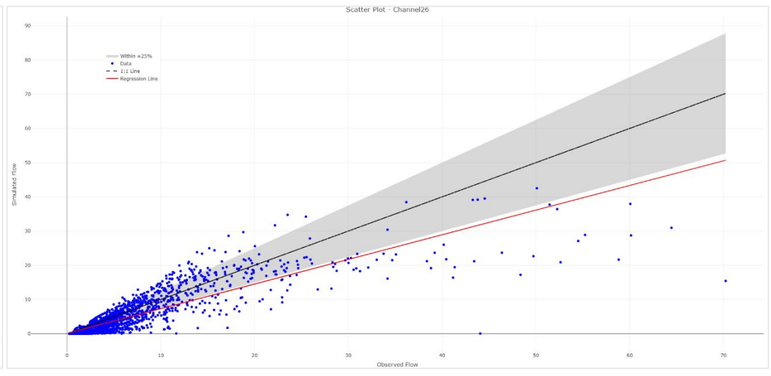




SC–ET Calibration (ET-only)







MC–QET Calibration (Q–ET)