Journal of Hydrology 664 (2026) 134527

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Towards water resilience: A multi-stage calibration framework for large-scale integrated surface–subsurface hydrological models

Leonardo Sandoval [a,b], Alberto Guadagnini [a], Laura Condon [b], Monica Riva [a,*]

[a] Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milan, Piazza L. Da Vinci 32, Milan, 20133, Italy
[b] Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ, USA

ARTICLE INFO

This manuscript was handled by Renato Morbidelli, Editor-in-Chief.

Keywords:
Surface–groundwater modeling
Po River basin
Parameter calibration
Sensitivity analysis
Water resource management
ParFlow-CLM

ABSTRACT

This study introduces a methodological framework for calibrating large-scale, high-fidelity integrated surface water–groundwater models with the aim of enhancing their reliability for water resource management. Our approach synergistically integrates ParFlow-CLM for simulating three-dimensional variably saturated flow, local sensitivity analysis to identify relevant model parameters, and Gaussian Process Regression surrogates for efficient multi-stage calibration against water table depth and river discharge observations. The framework is then exemplified through the analysis of the groundwater flow scenario associated with the Po River District (87,000 km^2) in northern Italy. The workflow yields the first robustly calibrated high-fidelity model at such spatial scale, embedding calibrated values of (*i*) hydraulic conductivities of the main geomaterials forming the internal architecture of the subsurface and (*ii*) Manning roughness coefficients of the major rivers in the domain. Our results highlight conductivity of clay as a dominant parameter driving groundwater table dynamics while channel roughness is the most important parameter for river flows. Our large-scale model calibration strategy offers a robust conceptual and computational environment for scenario analysis and sustainable water planning in the presence of climate- and anthropogenic-related pressures.

1. Introduction

Large-scale integrated, or partially integrated, modeling of surface and subsurface water systems has attracted growing interest in recent years (e.g., Scanlon et al., 2023; Naz et al., 2023). This is driven by the increasing need for reliable and robust physically based modeling tools to address concerns over potential impacts of climate and anthropogenic drivers on interconnected water systems.

In this context, a variety of regional, supra-regional, continental, and even global-scale modeling frameworks have been introduced to study surface water and groundwater (SW–GW) systems (e.g., Verkaik et al., 2024; Reinecke et al., 2024; Müller Schmied et al., 2021; Condon et al., 2021; Gleeson et al., 2021; Reinecke et al., 2019; de Graaf et al., 2017; Maxwell et al., 2015).

Despite these advances, larger-scale SW–GW modeling efforts remain constrained by several factors. A major limitation is the scarcity of high-resolution subsurface data. Moreover, the substantial computational demands of such numerical simulations often require simplified parameterizations of the subsurface environment, frequently treating groundwater as a static reservoir rather than a dynamic component of the hydrological cycle. In this broader context, and considering that groundwater is a vital resource for national economies worldwide, constitutes the main source of drinking water for up to two billion people (Gleeson et al., 2010), and supports more than half of the world's irrigated agriculture (Famiglietti, 2014), a detailed understanding of its dynamics is critical for effective water security planning.

Moreover, while robust parameter calibration techniques are well established for GW models at local or sub-regional scales (Janetti et al., 2021; Siena and Riva, 2020), their use in large-scale SW–GW models still remains limited. This is mainly due to the high dimensionality of the parameter space and the heavy computational burden associated with such models. As a result, SW–GW models frequently rely on highly simplified representations of the subsurface environment and related processes, frequently neglecting the unsaturated flow regime, subsurface spatial heterogeneity, and fully three-dimensional flow dynamics. Site-specific parameter calibration is rarely performed. Instead, model parameters are typically adjusted from literature values (e.g., Naz et al., 2023; Maxwell et al., 2015). These elements contribute to notably increase predictive uncertainty of integrated SW–GW models, thus limiting the reliability of model outputs for long-term forecasting and constraining their potential ability to inform sustainable water resource management at large spatial scales.

* Corresponding author.
E-mail address: monica.riva@polimi.it (M. Riva).

https://doi.org/10.1016/j.jhydrol.2025.134527
Received 28 August 2025; Received in revised form 9 October 2025; Accepted 29 October 2025
Available online 4 November 2025

In this study, we address these challenges by introducing a methodological framework for robust calibration of large-scale, high-fidelity, fully integrated SW–GW flow models. The key innovation of our approach lies in integrating diverse streams of research associated with the formulation of advanced tools that, to date, have not been jointly applied to investigate SW–GW dynamics at large (or supra-regional) scales. Our approach explicitly accounts for the three-dimensional subsurface heterogeneity and incorporates available (though often incomplete) system knowledge from available data.

The proposed framework integrates: (*i*) a fully coupled, physically based model of SW–GW processes with explicit representation of subsurface heterogeneity and three-dimensional features of unsaturated–saturated flow; (*ii*) a comprehensive sensitivity analysis to identify the most influential model parameters based on selected target variables; and (*iii*) a surrogate-model-assisted optimization strategy to sequentially calibrate key parameters using multiple available datasets. This integrated approach enables the simulation of current system dynamics and the evaluation of future scenarios, thereby strengthening our ability to support informed water management decisions under increasing environmental and societal pressures.

The proposed methodological approach is then employed on the largest groundwater system in Italy, the Po River watershed. The region serves as an ideal testbed due to its inherent heterogeneous nature (Previati et al., 2025; Manzoni et al., 2023) and complex hydrological challenges, including widespread groundwater contamination (Guadagnini et al., 2020; Colombani et al., 2016) and recurring issues of water scarcity (Carlson et al., 2025; Dall'Amico et al., 2025; Bozzola and Swanson, 2014). Notably, this study yields the first robustly calibrated, high fidelity, fully integrated surface water–groundwater large scale model. Attaining this goal is possible using the multi-component framework developed and implemented in this work.

The remainder of the paper is structured as follows. Section 2 describes the study area and summarizes previous modeling efforts. Section 3 details our modeling framework and parameter calibration methodology, together with the available data and the target variables of interest. Section 4 illustrates the main outcomes of our study. Conclusions and key remarks are provided in Section 5.

2. Study area

We apply our methodological framework to the Po River District, which spans an area of approximately 87,000 km^2 in northern Italy. The district, along with its principal watercourse, the Po River, plays a critical role in the country's economy, contributing to nearly 40% of the national gross domestic product. It provides water resources for approximately 24 million people (Istituto Nazionale di Statistica, 2020), supports irrigation across 1.6 million hectares of agricultural land (Montanari et al., 2023), that yields more than 35% of Italian agricultural production (Carlson et al., 2025), and sustains a broad spectrum of industrial activities. Given this strategic significance, a robust understanding of the district hydrological functioning is markedly relevant for preserving ecosystems integrity and ensuring socioeconomic resilience at both regional and national scales.

As shown in Fig. 1, the district is bounded by the Adige River to the northeast, the Adriatic Sea to the east, the Alps to the north and west, and the Apennines to the south. It spans 33 provinces across seven Italian regions (Aosta Valley, Lombardy, Piedmont, Emilia-Romagna, Liguria, Veneto, and Trentino-Alto Adige), the Swiss canton of Ticino, and parts of the French and Swiss Alps. The Po River basin constitutes the main watershed within this district, with a catchment area of approximately 71,000 km^2. The basin encompasses diverse climatic and geomorphological zones, ranging from high-altitude Alpine and Apennine regions (exceeding 3700 m above sea level and characterized by steep slopes, ≥75%, and sparse population densities, ≤1 inhabitant/km^2) to flat, low-lying plains with population density exceeding 2000 inhabitants/km^2 (Center For International Earth Science Information Network-CIESIN-Columbia University, 2018). Climatic conditions vary considerably across the basin due to differences in elevation and geographic position. Alpine areas exhibit distinct climatic zones associated with specific ecological characteristics and receive precipitation in the form of snow and rain (Agrawala, 2007; Elsasser and Bürki, 2002). Lowlands experience a continental temperate climate, with annual precipitation ranging from 600 to 900 mm (Grimm et al., 2003; Morgan, 1973). The foothill zone (termed Prealpi) experiences the highest annual precipitation (exceeding 2000 mm (Fratianni and Acquaotta, 2017; Morgan, 1973)). The Po River is characterized by an average discharge of about 1540 m^3/s (monitored at Pontelagoscuro flow rate station; see large blue triangle in Fig. 1) and receives inflow from more than 140 tributaries (Autorita di Bacino del Fiume Po, 2021).

The water resources of the Po River District face numerous compounding and interconnected pressures (Montanari et al., 2023). Climate change is significantly altering precipitation regimes and intensifying hydro-meteorological extremes, leading to seasonal shifts in runoff and expected reductions in water availability during spring and summer (Spinoni et al., 2018; Coppola et al., 2014; Beniston et al., 2007; Christensen and Christensen, 2003). At the same time, intensive agricultural, urban, and industrial activities contribute to elevated nutrient and pollutant loads, posing critical threats to surface and groundwater quality (Viaroli et al., 2018; Pieri et al., 2011; Pettine et al., 1996). Groundwater overexploitation further stresses the system, resulting in aquifer depletion and widespread land subsidence in sedimentary areas (Carlson et al., 2025; Xanke and Liesch, 2022; Stramondo et al., 2007). Coastal areas face additional threats from saltwater intrusion from the Adriatic Sea, a factor that undermines agricultural productivity and potable water supplies (Luo et al., 2024; Antonellini et al., 2008) while driving increased competition among users. Collectively, these pressures jeopardize water security, degrade ecosystem services, and undermine long-term socioeconomic resilience of the Po River District.

Several modeling efforts have previously targeted the Po River basin or its sub-catchments, each focusing on specific hydrological components and spatial–temporal scales. For example, Coppola et al. (2014) evaluate the impacts of climate change on the upper Po River discharge using a surface runoff model. Ravazzani et al. (2015) develop a hydrological model of the upper Po basin (~38,000 km^2) based on coupling a rainfall-runoff model with a two-dimensional saturated groundwater flow model to evaluate the effects of climate change as well as anthropogenic structures on water availability. In their approach, the groundwater system is simplified as two homogeneous storage regions (a shallow unconfined aquifer and a deeper confined aquifer) separated by an aquitard, SW–GW exchanges being allowed only along river channels (as driven through a Cauchy-type boundary condition). Vezzoli et al. (2015) evaluate climate-driven impacts on Po River discharge combining a distributed runoff model and a basin-scale water balance assessment (Liu and Todini, 2002). Cislaghi et al. (2020) apply a rainfall-runoff model across 63 mountainous sub-catchments of the Po watershed. More recently, Manzoni et al. (2024) conduct a steady-state GW modeling study of the Po River valley, simulating three-dimensional GW flow while accounting for uncertainties in the spatial distributions of geomaterials and incorporating a partial SW–GW coupling (through a Cauchy boundary condition). Notably, none of the above-mentioned studies develops a fully coupled, transient, robustly calibrated GW–SW model for the entire area. In our study, surface water and groundwater processes are jointly and simultaneously solved. This enables a physically consistent realistic representation of dynamic interactions between the two domains across all parts of the system that include rivers (and the land surface in general), while explicitly representing the natural heterogeneity and three-dimensional complexity of the groundwater system.

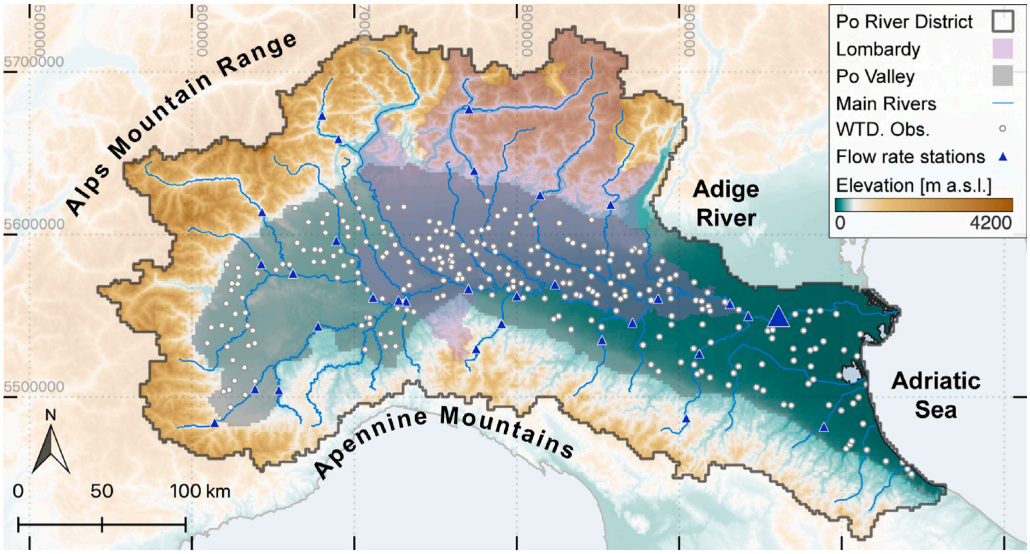

Fig. 1. Location and limits of the Po River District. Main rivers within the domain are displayed in blue. The Lombardy region and the Po River Valley are shaded in purple and gray, respectively. White circles indicate locations of water table depth (WTD) observations. Blue triangles denote flow rate measurement stations, the large blue triangle corresponding to the Pontelagoscuro station of the Po River. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Methodology

3.1. Modeling approach

Integrated surface–subsurface flow is simulated with ParFlow-CLM (Kollet and Maxwell, 2006), a physically-based numerical model that has demonstrated robust performance in reproducing hydrological dynamics across diverse hydroclimatic settings and spatial scales (Yang et al., 2023; Naz et al., 2023; Jefferson et al., 2017; Maxwell and Condon, 2016; Kollet and Maxwell, 2008). The model simulates three-dimensional variably saturated subsurface flow using the mixed form of the Richards' equation, i.e.,:

$$S_s S(\psi)\frac{\partial \psi}{\partial t} + \phi \frac{\partial S(\psi)}{\partial t} = \nabla \cdot \left[\underline{\underline{\mathbf{k}}} k_r(\psi) \nabla (\psi - z)\right] + q_s. \tag{1}$$

Here, S_s represents specific storage [L^{-1}], ψ is pressure head [L], t is time [T], ϕ is porosity [−], $\underline{\underline{\mathbf{k}}}$ is a tensor denoting saturated hydraulic conductivity [LT^{-1}], z is depth below terrain surface [L], and q_s is a source/sink term [T^{-1}]. The degree of saturation, $S(\psi)$ [−], and relative permeability, $k_r(\psi)$ [−], are evaluated as:

$$S(\psi) = \frac{S_{\mathrm{sat}} - S_{\mathrm{res}}}{\left[1 + (\alpha\psi)^n\right]^{1-1/n}} + S_{\mathrm{res}}, \tag{2}$$

$$k_r(\psi) = \frac{\left[1 - \frac{(\alpha\psi)^{n-1}}{(1+(\alpha\psi)^n)^{1-1/n}}\right]^2}{\left[1 + (\alpha\psi)^n\right]^{\frac{1-1/n}{2}}}, \tag{3}$$

where S_{sat} and S_{res} respectively denote saturated and residual water content [−], α [L^{-1}] and n [−] being soil parameters associated with the Van Genuchten model (Van Genuchten, 1980).

Surface flow is simulated using the depth-averaged Saint-Venant (or shallow water) equations, i.e.

$$\frac{\partial \psi_s}{\partial t} + \nabla \cdot (\mathbf{v}\psi_s) = q_r, \tag{4}$$

where $\mathbf{v} = [v_x, v_y]$ is the two-dimensional depth-averaged velocity vector [LT^{-1}], ψ_s is the water depth (or surface ponding depth), and q_r is a source/sink rate term [LT^{-1}] (corresponding, e.g., to rainfall rate). Eqs. (1) and (4) are coupled by enforcing continuity of pressure and flux at the ground surface. Velocity components are evaluated upon resting on the kinematic wave approximation:

$$v_i = \frac{\sqrt{|S_i|}}{\eta}\psi_s^{2/3}\frac{S_i}{|S_i|}, \text{ with } i = x, y, \tag{5}$$

where S_x [−] and S_y [−] represent topographic slopes in the west–east and south–north directions, respectively, and η is the Manning surface roughness coefficient [$\mathrm{TL}^{-1/3}$].

Land surface water and energy fluxes are simulated using a modified version of the Common Land Model - CLM (Dai et al., 2003), which is fully integrated into ParFlow. The land surface is structured across discrete units (termed tiles) where key fluxes, such as evaporation, transpiration, and infiltration are evaluated. Bare ground evaporation is determined from specific humidity, air density, and resistance terms, while transpiration is evaluated as a function of leaf and stem area indices, air density, and boundary layer resistance (Jefferson et al., 2017). Snow water equivalent is modeled using thermal, vegetation, canopy, and snow age processes, which determine the amount of precipitation falling as snow. Temporal snow dynamics are simulated through albedo decay, snow compaction, sublimation, and melt processes (Ryken et al., 2020). Detailed formulations of the governing equations and the integration of surface–subsurface processes in ParFlow-CLM, can be found in Kollet and Maxwell (2006) and Maxwell et al. (2024).

3.2. Model setup

3.2.1. Geometry, boundary conditions and hydrogeological parametrization

The numerical model domain (see Fig. 1) extends for 530 km and 330 km in the east–west and south–north direction, respectively. A uniform horizontal grid resolution of 2 km × 2 km is employed. The domain comprises active and inactive grid cells, with the active cells encompassing the entire Po River District. Vertically, the domain extends to a depth of 225 m, and is discretized into six layers of variable thicknesses: 170 m, 40 m, 13 m, 1.4 m, 0.45 m, and 0.15 m (from bottom to top). This grid design provides enhanced resolution of near-surface flow dynamics. A terrain-following grid is employed (Maxwell, 2013), resulting in a total of 130,536 computational cells (21,756 active lateral cells across six subsurface layers). A Dirichlet boundary condition is imposed at the boundary with the Adriatic Sea, with the hydraulic head set to sea level. Zero-flow boundary conditions are imposed at the remaining lateral boundaries (coinciding with the mountain crests) and at the bottom of the domain.

The spatial distribution of subsurface facies is obtained using two complementary datasets, i.e., (i) the three-dimensional probabilistic hydrostratigraphic reconstruction developed by Manzoni et al. (2023) and (ii) the SoilGrids dataset (Poggio et al., 2021).

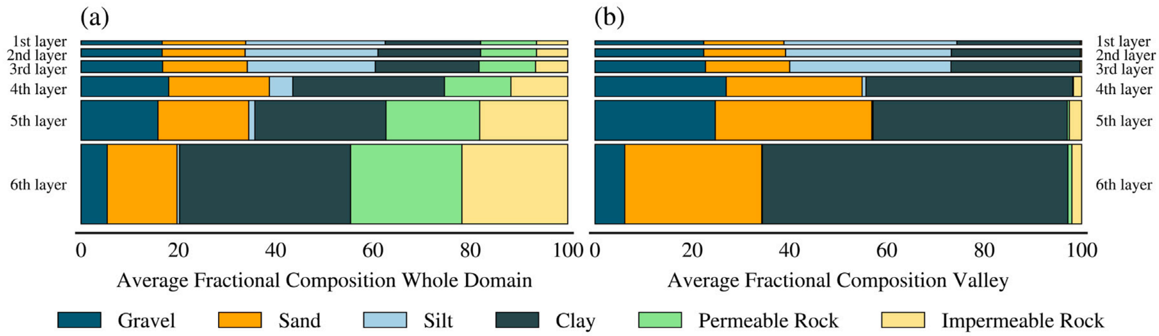

Fig. 2. Average fractional composition of geomaterial types across each model layer for (a) the entire domain and (b) the valley region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The facies reconstruction of Manzoni et al. (2023) is based on a neural network algorithm to yield estimates of the spatial distribution and associated uncertainty of six main geomaterial types (i.e., gravel, sand, silt, clay, permeable rock, and impermeable rock) at a resolution of 1 km × 1 km × 1 m. For each voxel, these authors evaluate the probability of occurrence of each geomaterial. To estimate the way a given geomaterial contributes to the overall composition of each cell of our numerical model, we then evaluate the mean fractional contribution of each geomaterial type associated with the voxels residing therein. Similarly, fractional contributions of clay, silt, and sand for the upper three layers of the domain (corresponding to depths from 0 m to 2 m) are estimated using data from the SoilGrids dataset (Poggio et al., 2021). Fig. 2 depicts the resulting average fractional contribution of each geomaterial across all layers of the numerical model, with panel (a) representing the full domain and panel (b) focusing on the valley area (highlighted in dark gray in Fig. 1). A depth-dependent trend is evident, whereby deeper layers are characterized by higher fractions of less permeable geomaterials.

Values of parameters α, n, and ϕ for each cell in the numerical model are assessed as weighted linear combinations of the corresponding values assigned to each geomaterial type, with weights corresponding to the fractional contribution of each geomaterial within the cell. Similarly, the equivalent saturated hydraulic conductivity (hereinafter referred to as conductivity, for simplicity) along the vertical direction for each model cell is evaluated through a weighted harmonic mean of the conductivity values assigned to each geomaterial type, with weights corresponding to their fractional contributions within the cell. This approach is based on a conceptual picture according to which most subsurface flow takes place in the valley, where geomaterials exhibit a predominantly layered structure. In such settings, the harmonic mean is typically considered to provide an appropriate operational estimate of an equivalent vertical conductivity (Sanchez-Vila et al., 2006; Terzaghi et al., 1996). An anisotropy ratio of 100 is applied uniformly across all model cells (i.e., horizontal conductivity is assumed to be 100 times higher than vertical conductivity). This value reflects conditions that are typical of sedimentary environments. We also find it to be generally consistent with the corresponding ratio between the above mentioned estimated vertical equivalent conductivity and an equivalent horizontal conductivity value resulting from a weighted linear mean of geomaterial conductivities.

A uniform value across the domain is considered for specific storage. This is in line with a conceptual picture according to which propagation of hydraulic head perturbations (governed by the ratio between transmissivity and storage) reflects only the spatial heterogeneity of K. Although this streamlined picture neglects possible spatial variability in the specific storage of the system, it enables us to clearly isolate and analyze the effects of conductivity heterogeneity on groundwater dynamics while maintaining model tractability. Moreover, since calibration of hydraulic conductivity is based on steady-state simulations (see Section 3.3.3), specific storage does not influence the ensuing water table depth distribution rendered by the model once it has reached steady-state.

Finally, two distinct values are specified for the Manning roughness coefficients, denoted as η_1 and η_2 for land and channel cells, respectively.

3.2.2. Input data

The Shuttle Elevation Derivatives at multiple Scales (HydroSHEDS) digital elevation model (Lehner et al., 2021) (See Fig. 3a) is processed using the PriorityFlow package (Condon and Maxwell, 2019) to obtain west–east (S_x) and south–north (S_y) topographic slopes that are required in Eq. (5). HydroSHEDS provides a 30 arcseconds (approximately corresponding to 900 m at the latitude of our studythe domain area) resolution DEM.

Land cover data are obtained from the MODIS MCD12Q1 product (Friedl and Sulla-Menashe, 2019), using the IGBP classification. Annual maps from 2009 to 2018 with a spatial resolution of 500 m are gathered for this study. Land cover in the district shows limited interannual variability. Fig. 3b illustrates a simplified version (lumped land covers) of the land cover classification for 2009, where similar classes are grouped for clarity. Note that all 17 IGBP classes are retained in our numerical model.

Meteorological forcing variables (i.e., precipitation, wind speed (zonal and meridional components), near-surface temperature, shortwave and longwave radiation, surface pressure, and specific humidity) are obtained from the COSMO-REA6 reanalysis dataset (Bollmeyer et al., 2015) at hourly resolution for the period 2009–2018. The native spatial resolution of COSMO-REA6 is 0.055° (corresponding to approximately 6.1 km at the latitude of the model domain). For illustration purposes, Fig. 3c and d depict the spatial distribution of mean annual precipitation and temperature fields over the study period.

Evapotranspiration data are obtained from the fifth version of the Global Land Surface Satellite (GLASS) evapotranspiration product (Xie et al., 2022), available at an 8-day temporal resolution and 1 km spatial resolution for the period 2009–2018. Fig. 3e illustrates the average annual evapotranspiration across the domain. This dataset is employed to estimate potential recharge under steady-state conditions, as further discussed in Section 3.3.3.

All datasets are reprojected and resampled to a common 2 km × 2 km spatial resolution, to ensure consistency with the numerical model resolution. Forcing data are subject to bilinear interpolation using the climate data operator (Schulzweida, 2023). Otherwise, the digital elevation model and the evapotranspiration data are resampled using GIS tools taking the mean value across the tiles overlying each cell of the numerical model, while the land cover data being resampled using GIS tools taking the most frequent land surface category of such tiles.

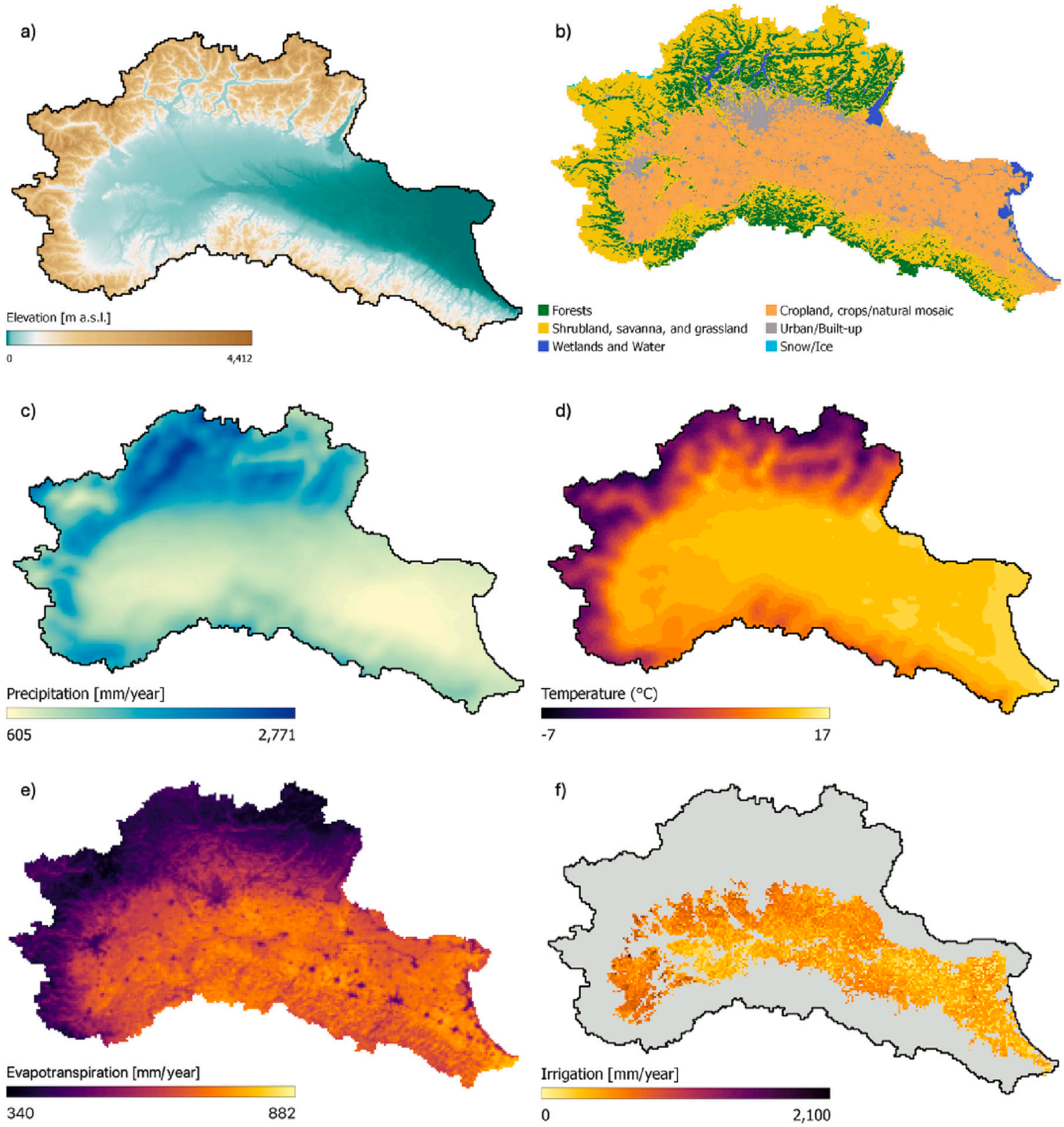

Fig. 3. Input datasets used to parameterize and define boundary conditions of the numerical model: (a) elevation (from the HydroSHEDS digital elevation model); (b) land cover classification (for 2009 from the MODIS MCD12Q1 product, using the IGBP scheme); (c) mean annual precipitation [mm yr^{-1}], and (d) mean annual temperature [°C], from the COSMO-REA6 reanalysis (2009–2018); (e) mean annual evapotranspiration [mm yr^{-1}], from the GLASS ET dataset; (f) mean annual irrigation [mm yr^{-1}] from Dari et al. (2023). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Our model accounts for groundwater abstraction distributed across domestic, agricultural, and industrial sectors, amounting to a total of 6034.6 Mm^3 yr^{-1} .

Domestic water use is estimated using two approaches, depending on data availability. In the Lombardy region (purple-shaded area in Fig. 1), values are taken directly from local water authority reports. For the remaining regions, consumption is estimated as the product of per-capita groundwater use (ranging from 5 to 131 m^3 yr^{-1}, derived from Lombardy data and assigned to the remaining regions based on geographical similarity to provinces in Lombardy) and population statistics associated with each province from ISTAT (Italian National Institute of Statistics) 2024. This approach yields a total domestic groundwater use of approximately 1910 Mm^3 yr^{-1}, which aligns well with national estimates of around 2000 Mm^3 yr^{-1} (Autorita di Bacino del Fiume Po, 2012). Well location and screen depth data are available for the Lombardy region. Elsewhere, wells are assumed to be uniformly distributed across the plains of each province, with filter depths ranging from 150 m to 200 m below ground surface. This assumption is consistent with standard engineering practice for drinking water wells, which typically target deeper aquifers to minimize contamination risks.

Agricultural groundwater use for irrigation is estimated using regional datasets from Dari et al. (2023), which provide weekly data at 1 km spatial resolution for the period 2016–2019. Fig. 3f depicts the annual average irrigation rate derived from this dataset. The estimated total withdrawal for irrigation use is 16,307 Mm^3 yr^{-1}, in close agreement with the official figure of 16,500 Mm^3 yr^{-1} reported by National authorities (Autorita di Bacino del Fiume Po, 2012). An amount corresponding to 17% of this quantity (i.e., 2805 Mm^3 yr^{-1}) is attributed to groundwater abstractions (Autorita di Bacino del Fiume Po, 2012).

Industrial water use is quantified from the Po River Watershed Authority (Autorita di Bacino del Fiume Po, 2012) data. The latter reports a total annual abstraction of 1537 $\mathrm{Mm^3\ yr^{-1}}$ for industrial activities, 80% of which (i.e., 1229.6 $\mathrm{Mm^3\ yr^{-1}}$) is sourced from groundwater.

3.3. Model calibration

This section outlines the procedure employed to estimate hydraulic parameters through calibration of the numerical model. First, we describe the methodology used to identify the parameters with the highest influence on model outputs and that are then considered in the context of model calibration (Section 3.3.1). We then present the calibration dataset (Section 3.3.2) and the overall calibration strategy, including the optimization algorithm and the use of (reduced-complexity) surrogate models to ensure computational feasibility (Section 3.3.3).

3.3.1. Identification of the most influential model parameters

We rely on a local sensitivity analysis approach to quantify the effect of individual model parameters on selected quantities of interest. The sensitivity analysis is performed according to a one-at-a-time approach (Saltelli, 1999). The model $\mathcal{M}$ is first executed using a baseline parameter set $\boldsymbol{\theta}_o = (k_i, \phi_i, \alpha_i, n_i, S_s, \eta_1, \eta_2)$, with $i = 1, \ldots, 6$ (i = 1: Gravel, 2: Sand, 3: Silt, 4: Clay, 5: Permeable rock, and 6: Impermeable rock). The resulting output is denoted as $\mathbf{y_o} = \mathcal{M}(\boldsymbol{\theta}_o)$. Each parameter is then perturbed one-at-a-time (i.e., with all other parameters held constant), first by increasing (yielding $\mathbf{y_{1^+}} = \mathcal{M}(\boldsymbol{\theta}_{1^+})$) and then by decreasing (to obtain $\mathbf{y_{1^-}} = \mathcal{M}(\boldsymbol{\theta}_{1^-})$) its value. The mean of the absolute differences between the perturbed and baseline model outputs is evaluated. This procedure is repeated for all model parameters. Among the latter, parameters associated with the largest average variations are identified as the most influential ones and adjusted during the calibration process. The remaining parameters are fixed at their baseline value.

Table 1 lists the initial values (corresponding to the baseline parameter set, $\boldsymbol{\theta}_o$) of the model parameters used in this study. These values are informed by previous calibration efforts and analyses of hydraulic properties in the area (Previati et al., 2025; Manzoni et al., 2024), standard reference tables in hydrogeology (Bear, 2007; Bear et al., 1968), and parameter values assigned to similar geomaterial types in other large-scale modeling studies (Maxwell and Condon, 2016).

Each parameter type is perturbed according to its expected range of variability (see Table 2). The initial values of hydraulic conductivity and specific storage are increased and decreased by one order of magnitude to account for their typically high variability in natural systems (see Table 2.4.1 of Bear and Cheng (2010) and Table 2–14 of Batu (1998)). The initial values of the Manning roughness coefficients is also perturbed by one order of magnitude. This is consistent with the observation that the spatial scale of our model (2 km horizontal resolution) implies that each grid cell represents a mixture of multiple surface types, resulting in high uncertainty in the effective Manning coefficient. In contrast, porosity and Van Genuchten parameters typically exhibit smaller ranges of variability (see Table 2.4.1 of Bear and Cheng (2010), and Table S1 of Maxwell and Condon (2016)). Hence, we subject the initial value of these parameters to a milder perturbation compared to the other parameters. Specifically, we perturb the initial value of ϕ by ± 0.1, α by $\pm 0.5\ \mathrm{m^{-1}}$, and n by ± 0.5.

3.3.2. Calibration dataset

Fig. 1 depicts the locations of the piezometers (white dots) where water table depth (WTD) data are available. The dataset, encompassing the period 2012–2018, is sourced from local authorities of the three largest regions within the model domain (i.e., Lombardy, Piedmont, and Emilia-Romagna) and comprises measurements from 285 observation wells. Data are processed following the methodology described in Manzoni et al. (2024). Of the original dataset, 41 wells are located on channel cells of our model and are excluded from the analysis.

In addition, daily discharge data from 31 gauging stations (blue triangles in Fig. 1) are obtained from the EStreams dataset (do Nascimento et al., 2024). Stations are selected on the basis of three criteria: (i) a drainage area of at least 500 $\mathrm{km^2}$, consistent with the large-scale of our model and threshold used in comparable studies (Naz et al., 2023); (ii) agreement within a 20% tolerance between modeled drainage areas and the corresponding area reported in do Nascimento et al. (2024); and (iii) time series completeness, with no more than 30% missing data during 2009–2018.

3.3.3. Calibration strategy and techniques

Estimation of (i) conductivity for selected geomaterials and (ii) Manning roughness coefficient for channel cells is performed according to two sequential phases. Other model parameters are excluded from calibration based on the results of the sensitivity analysis (see Section 4.1).

In the first phase, geomaterial conductivities are calibrated by comparing time-averaged WTD observations from the 244 piezometers described in Section 3.3.2 against their model-simulated counterparts under constant forcing. The model is forced with a spatially distributed (time-averaged) potential recharge field. The latter is evaluated as the mean over 2009–2018, combining precipitation, irrigation and evapotranspiration data. Although the irrigation dataset covers only 2016–2019, it is assumed to be representative of the full simulation period. Simulations proceed until the maximum annual variation in WTD is less than 1 mm in cells containing the piezometers. At this point, annual storage changes are below 1% of potential recharge, consistent with standard steady-state convergence criteria (Yang et al., 2023). It is noted that calibration of conductivity values using steady-state conditions is a broadly adopted and computationally efficient approach that captures the dominant hydraulic behavior of the system (Manzoni et al., 2024; Anderson et al., 2015).

Estimation of conductivity values is performed upon minimizing the root mean squared error (RMSE) between model results and available observations. We recall that when model errors (i.e., the differences between simulated and observed values) are independent and normally distributed with zero mean and measurement error variance is fixed, minimization of the RMSE leads to Maximum Likelihood estimates of parameter values (Carrera and Neuman, 1986).

Implementation of the model calibration approach requires multiple evaluations of model outputs across different combinations of model parameters $\boldsymbol{\theta}$. For large scale modeling, as in this study, this becomes computationally unfeasible. To overcome this limitation, we perform model calibration with Gaussian Process Regression (GPR) surrogates of the full numerical model. GPR constitutes a general Bayesian regression framework capable of approximating complex functional relationships without requiring explicit knowledge of the underlying generative process between model inputs (μ) and outputs (ν).

The estimate of a quantity of interest (ν^*) and its associated uncertainty ($\sigma^2_{\nu^*}$) at a new unsampled location (μ^*) are evaluated as:

$$\begin{aligned} \nu^* &= \underline{\mathbf{m}}^{\mathrm{T}} \underline{\underline{\mathbf{M}}}^{-1} \underline{\boldsymbol{\nu}}, \\ \sigma^2_{\nu^*} &= \operatorname{cov}\left[\nu(\mu^*), \nu(\mu^*)\right] - \underline{\mathbf{m}}^{\mathrm{T}} \underline{\underline{\mathbf{M}}}^{-1} \underline{\mathbf{m}}, \end{aligned} \tag{6}$$

where $\underline{\mathbf{m}} = \left\{ \operatorname{cov}\left[\nu(\mu^*), \nu(\mu_1)\right], \ldots, \operatorname{cov}\left[\nu(\mu^*), \nu(\mu_{N_{obs}})\right] \right\}$ is the vector of covariances between the output at the new unsampled point, $\nu(\mu^*)$, and the output at all N_{obs} observation points; $\underline{\underline{\mathbf{M}}}$ is a $N_{obs} \times N_{obs}$ matrix, whose entry $M_{i,j}$ is $\operatorname{cov}\left[\nu(\mu_i), \nu(\mu_j)\right]$; and $\underline{\boldsymbol{\nu}} = \left[\nu(\mu_1), \nu(\mu_2), \ldots, \nu(\mu_{N_{obs}})\right]$ is the vector containing the value of the quantity of interest at the N_{obs} observation points. Vector $\underline{\mathbf{m}}$ and matrix $\underline{\underline{\mathbf{M}}}$ are constructed using the specific covariance function defined for GPR. In this study, we employ a

Table 1
Initial values of the model parameters, θ_o.

Variable [Units]	Initial Value	Source of information for initial parameter value
k_1 [m h^{-1}]	3.60×10^{1}	Calibrated by Manzoni et al. (2024)
k_2 [m h^{-1}]	1.14	In range of Bear and Cheng (2010)
k_3 [m h^{-1}]	1.14×10^{-3}	In range of Bear and Cheng (2010)
k_4 [m h^{-1}]	3.60×10^{-3}	In range of Previati et al. (2025)
k_5 [m h^{-1}]	3.60×10^{-5}	Order of magnitude from Manzoni et al. (2024)
k_6 [m h^{-1}]	3.60×10^{-7}	Order of magnitude from Manzoni et al. (2024)
ϕ_1 [–]	0.35	In range of Bear and Cheng (2010)
ϕ_2 [–]	0.38	
ϕ_3 [–]	0.49	
ϕ_4 [–]	0.46	
ϕ_5 [–]	0.30	
ϕ_6 [–]	0.33	
α_1 [m^{-1}]	3.55	Similar material from Maxwell and Condon (2016)
α_2 [m^{-1}]	3.55	
α_3 [m^{-1}]	0.66	
α_4 [m^{-1}]	1.51	
α_5 [m^{-1}]	1.00	
α_6 [m^{-1}]	1.00	
n_1 [–]	4.16	Similar material from Maxwell and Condon (2016)
n_2 [–]	4.16	
n_3 [–]	2.66	
n_4 [–]	2.26	
n_5 [–]	3.00	
n_6 [–]	3.00	
η_1 [h m$^{-1/3}$]	1.00×10^{-5}	Mid-range value across all surface materials from Chow (1959)
η_2 [h m$^{-1/3}$]	1.00×10^{-5}	Mid-range value across all surface materials from Chow (1959)
S_s [m^{-1}]	1.60×10^{-4}	Mid-range value across all geomaterials from Batu (1998)

Gaussian covariance function, also known as radial basis function (RBF) kernel, defined as:

$$\mathrm{cov}\left[\nu(\mu'), \nu(\mu'')\right] = \sigma_f^2 \exp\left(-\frac{|\mu' - \mu''|^2}{2l^2}\right), \tag{7}$$

where σ_f^2 is the variance associated with the target model output of interest, l is a length scale parameter that controls the rate at which correlation decays with distance in the model input space, and μ' and μ'' correspond to two points in the input space.

Construction of GPR surrogates requires the estimation of the hyperparameters σ_f^2 and l. These are evaluated through an optimization algorithm that minimizes the RMSE between outputs from the full numerical model and the corresponding results rendered through the surrogate model. In general, the accuracy of the surrogate increases and the uncertainty associated with model outputs decreases as the number of full model simulations used for training the surrogate increases. In this study, a total of 45 numerical simulations are performed to train the surrogate model.

In the second phase of the calibration process, the Manning roughness coefficient for channel cells is estimated by minimizing the RMSE between daily river flow observations at the gauging stations described in Section 3.3.2 and their simulated counterparts. This phase of model calibration is performed under transient conditions at hourly resolution, forced with time-varying meteorological inputs from the COSMO-REA6 dataset for the period 2009–2018.

We implement a two-step spin-up procedure to minimize the influence of initial conditions on calibration results and to properly initialize the model. In the first step, the model is run using conductivity values calibrated in the first phase and a fixed Manning coefficient (i.e., the baseline value). Meteorological forcing is kept constant in time, matching the conditions used for conductivity calibration. The simulation proceeds until steady-state conditions are attained, these being defined as the state at which changes in total water storage and valley-average water table depth (WTD_V) fall below 1% of potential recharge and 1 mm/year, respectively. In the second step, dynamic spin-ups are performed independently by varying η_2 and using meteorological forcing from the COSMO-REA6 dataset for the year 2009, which represents a climatologically typical year for the study. Each spin-up continues until annual changes in (i) subsurface storage, (ii) surface storage, and (iii) snow water equivalent are below 1% of their respective aggregated values. This threshold value is consistent with criteria adopted in comparable studies (Seck et al., 2015). Following spin-up, the full 2009–2018 simulation period is executed.

3.4. Computational resources

The simulations used in this study are executed on 40 processors by partitioning the model domain into 8 × 5 subdomains along the x and y directions, respectively. Computations are performed (non-simultaneously) on three high-performance computing (HPC) systems: (i) the CFDHub@Polimi, an interdepartmental facility at Politecnico di Milano, using 40 cores of an Intel® Xeon® Gold 6148 CPU operating at 2.40 GHz; (ii) the Galileo100 system at CINECA, employing nodes with dual Intel CascadeLake 8276 CPUs (24 cores each, 2.4 GHz); and (iii) the LEONARDO HPC system at CINECA, using nodes equipped with dual Intel Xeon Platinum 8480 CPUs (56 cores, 2.0 GHz). Simulating one hydrologic year requires (on average) approximately 200 core-hours.

4. Results and discussion

4.1. Local sensitivity analysis

Fig. 4 illustrates the relative importance of model parameters (listed in Table 1) on key quantities of interest of our study, namely: water table depth in the valley (WTD_V), flow rate at the Pontelagoscuro river station (Q_P), subsurface water storage (Sub_S), and surface water storage (Sur_S). The relative importance of a parameter is evaluated as its average induced variation (as described in Section 3.3.1), normalized, for illustration purposes, by the sum of all parameter-induced variations. This analysis is performed by imposing a time-averaged potential recharge field (as described in Section 3.3.3) and stopping the simulations once steady-state conditions are attained.

Fig. 4 suggests that WTD_V is primarily influenced by the hydraulic conductivities of the (geologic) facies. This finding is consistent with Engdahl (2024), who report that conductivity plays a primary role

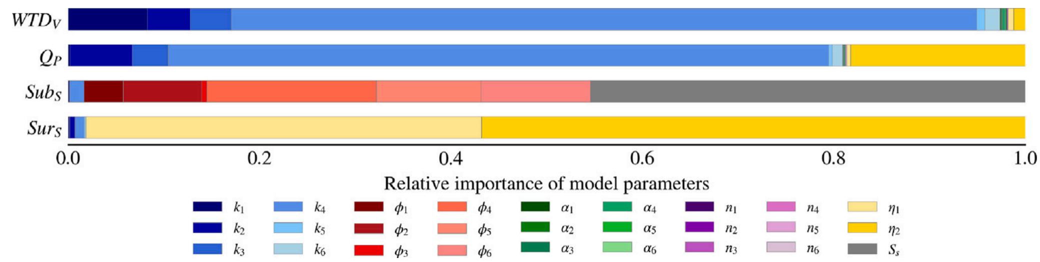

Fig. 4. Results of the sensitivity analysis, showing the relative importance of model parameters on key quantities of interest: water table depth in the valley (WTD_V), Po River discharge at Pontelagoscuro (Q_P), subsurface water storage (Sub_S), and surface water storage (Sur_S). Each bar corresponds to a given quantity of interest; the length of each segment indicates the (normalized) contribution of each model parameter. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in shaping the steady-state WTD behavior of the Fourth of July Creek in Idaho (U.S.). In the Po Valley, clay conductivity (k_4) emerges as the most influential parameter for WTD_V, reflecting the broad presence of clay in the area (see Fig. 2). Gravel (k_1), sand (k_2), and silt (k_3) conductivities also play a notable role, whereas the impact of permeable and impermeable rocks is negligible. Table 2 lists the percentage variations in each quantity of interest, defined as the difference between the perturbed-simulation result and the baseline value, normalized by the baseline value. These results suggest that increasing gravel and sand conductivity contribute to raise WTD_V, while the opposite can be observed upon increasing silt conductivity, due to its presence mostly in the upper two meters of the system. Clay conductivity exerts a more complex effect: either increasing or decreasing k_4 can produce localized rises or drops in WTD_V, an effect which cannot be fully captured by examining the overall WTD in the plain.

The flow rate of the Po River, Q_P, is also primarily governed by the facies conductivities, particularly by k_4. Increasing clay, sand, or silt conductivity reduces Q_P, reflecting the observation that higher conductivities facilitate infiltration and groundwater flow, thereby decreasing surface flow, as previously reported also by Foster and Maxwell (2019). Manning roughness coefficient for channel cells (η_2) also exerts substantial influence on Q_P, increased values of η_2 yielding a reduction of flow rates and vice versa. It is otherwise noted that the overall impact of parameter variations on Q_P is modest, the largest changes remaining below 2% of the average flow rate (see Table 2). This aligns with results from Engdahl (2024), who find that variability in river flow rate is approximately two orders of magnitude smaller than the mean flow rate when varying conductivity values.

Subsurface storage (Sub_S) is chiefly influenced by porosity and specific storage parameters, with impacts that are approximately 100 to 1000 times greater than those of other parameters. From a phenomenological standpoint, this is expected, because higher porosity and specific storage allow for greater water accumulation for a given WTD value. The direction of this variation also reconciles with a typical conceptual picture of a subsurface system, where the value of subsurface storage increases with porosity and specific storage values (See Table 2). Clay porosity is seen to have the largest effect on Sub_S, consistent with its abundance across the domain, especially in the valley. Porosity of permeable and impermeable rock also have a significant impact on Sub_S, this effect being related to their high abundance across the mountains of the domain.

Note that specific storage (defined as $S_s = \gamma[\phi\beta + (1-\phi)\alpha_s]$, where γ is the specific weight of water, β is water compressibility, α_s is soil compressibility, and ϕ is porosity) is a function of porosity (Bear and Cheng, 2010). While γ and β are quantities associated with relatively low degree of uncertainty under typical aquifer conditions, porosity and soil compressibility can be much more uncertain in the system. Hence, uncertainty associated with S_s should be interpreted as a result of joint uncertainties of porosity and soil compressibility. The relative importance of S_s in Fig. 4 should then be considered to be at least in part due to porosity uncertainty.

Surface storage (Sur_S) is primarily influenced by Manning roughness coefficients (η_1 and η_2). Higher roughness values provide increased resistance to surface flow, thus prolonging water retention in the system and raising surface storage as quantified in Table 2. Channel roughness (η_2) has a slightly greater impact than land roughness (η_1), as most surface flow takes place in channels, despite their total spatial extent being significantly smaller compared to the overall extent of the domain surface.

Overall, the sensitivity analysis highlights that the available dataset, comprising water table and discharge data (see Section 3.3.2) is mainly affected by granular material conductivities (k_1, k_2, k_3, and k_4) and by the Manning roughness coefficient of channels. Specifically, conductivities k_1, k_2, k_3, and k_4 strongly control WTD_V and Q_P, the latter being also influenced by η_2. Therefore, the calibration efforts presented in the following sections will focus on these parameters. It is worth noting that performing the sensitivity analysis using a time-averaged potential recharge field may mask sensitivities associated with dynamic hydrological responses. Consequently, the results presented here should be interpreted in light of this constrain.

4.2. Calibration of facies conductivities

The first phase of model calibration focuses on estimating optimal values for k_1, k_2, k_3, and k_4. A stagewise calibration strategy is herein employed, where three parameters are fixed in each stage while calibrating the remaining one. This process is repeated for all parameters until the selected error metric (i.e., RMSE; see Section 3.3.3) stabilizes. As shown in Table 3 the process requires eights stages. Calibration begins with k_4, as sensitivity analysis (Section 4.1) identifies it as having the greatest influence on WTD_V.

Columns 2–5 of Table 3 list the fixed parameters value and the support ranges (in square brackets) for the parameter being calibrated at each stage. The number of parameter value instances used for ParFlow simulations (indicated in parentheses after each support range) is sampled evenly in log space. Results of these simulations are employed to construct GPR surrogates of WTD at pixels where observations are available for calibration. Between 5 and 7 sampling points are used per stage, additional points being added as needed to improve the GPR accuracy. All simulations are run until achieving steady-state conditions, as described in Section 3.3.3. This ensures that surrogate quality is not affected by simulations that have not yet achieved steady-state conditions.

GPR surrogate models are then used to calibrate facies conductivities. For each stage listed in Table 3, 1000 conductivity values are sampled evenly (in log space) over the support range and passed to GPR models to compute WTD at observation pixels. The RMSE value between simulated and observed WTD (see dashed curves in Fig. 5) is computed as a function of the facies conductivity being calibrated. The conductivity value minimizing RMSE is selected and carried forward in the next stage. The optimal conductivity values for each stage are listed in column 6 of Table 3. Fig. 5 also shows values of

Table 2
Perturbed parameters, corresponding baseline values, and resulting percentage change in the quantities of interest.

Perturbed Parameter [units]	Baseline value	Parameter value	Variation Sub_S [%]	Variation Sur_S [%]	Variation WTD_V [%]	Variation Q_P [%]
k_1 [m h^{-1}]	3.60×10^{1}	3.60×10^{2}	−0.02	−0.22	2.50	0.00
		3.60	0.07	0.18	−6.38	0.00
k_2 [m h^{-1}]	1.14	1.14×10^{1}	−0.03	−2.13	2.72	−0.07
		1.14×10^{-1}	0.01	0.27	−2.17	0.18
k_3 [m h^{-1}]	1.14×10^{-3}	1.14×10^{-2}	−0.01	−0.17	−0.25	−0.03
		1.14×10^{-4}	−0.03	0.28	4.42	0.11
k_4 [m h^{-1}]	3.60×10^{-3}	3.60×10^{-2}	−0.24	−2.08	15.24	−2.14
		3.60×10^{-4}	−1.25	1.23	69.97	0.52
k_5 [m h^{-1}]	3.60×10^{-5}	3.60×10^{-4}	−0.09	−0.27	−0.47	−0.01
		3.60×10^{-6}	0.02	0.20	0.50	0.01
k_6 [m h^{-1}]	3.60×10^{-7}	3.60×10^{-6}	−0.07	−0.24	1.45	0.04
		3.60×10^{-8}	0.00	0.12	−0.20	−0.01
ϕ_1 [–]	0.35	0.45	1.94	0.00	0.00	0.00
		0.25	−1.94	0.00	0.00	0.00
ϕ_2 [–]	0.38	0.48	3.94	0.00	0.00	0.00
		0.28	−3.94	0.00	0.00	0.00
ϕ_3 [–]	0.49	0.59	0.26	0.00	0.00	0.00
		0.39	−0.26	0.00	0.00	0.00
ϕ_4 [–]	0.46	0.56	8.45	0.00	0.00	0.00
		0.36	−8.46	0.00	0.00	0.00
ϕ_5 [–]	0.30	0.40	5.19	0.00	0.00	0.00
		0.20	−5.19	0.00	0.00	0.00
ϕ_6 [–]	0.33	0.43	5.49	0.00	0.00	0.00
		0.23	−5.49	0.00	0.00	0.00
α_1 [m^{-1}]	3.55	4.00	−0.01	0.00	−0.02	0.00
		3.00	0.00	0.00	0.01	0.00
α_2 [m^{-1}]	3.55	4.00	−0.01	0.00	0.00	0.00
		3.00	0.00	0.00	−0.02	0.00
α_3 [m^{-1}]	0.66	1.16	−0.01	0.00	0.00	0.00
		0.16	0.00	0.00	−0.03	0.00
α_4 [m^{-1}]	1.51	2.00	−0.02	0.00	0.10	0.00
		1.00	0.00	0.00	−0.16	0.00
α_5 [m^{-1}]	1.00	1.50	−0.03	0.00	0.00	0.00
		0.50	0.02	0.00	−0.02	0.00
α_6 [m^{-1}]	1.00	1.50	−0.06	0.00	0.00	0.00
		0.50	0.03	0.00	−0.01	0.00
n_1 [–]	4.16	4.66	−0.01	0.00	−0.02	0.00
		3.66	0.01	0.00	0.01	0.00
n_2 [–]	4.16	4.66	0.00	0.00	0.00	0.00
		3.66	0.00	0.00	−0.01	0.00
n_3 [–]	2.66	3.16	−0.01	0.00	−0.01	0.00
		2.16	0.00	0.00	−0.01	0.00
n_4 [–]	2.26	2.76	−0.02	0.00	0.01	0.00
		1.76	0.02	0.00	−0.02	0.00
n_5 [–]	3.00	3.50	0.00	0.00	0.00	0.00
		2.50	−0.01	0.00	−0.01	0.00
n_6 [–]	3.00	3.50	0.01	0.00	−0.01	0.00
		2.50	−0.01	0.00	0.00	0.00
η_1 [h m$^{-1/3}$]	1.00×10^{-5}	1.00×10^{-4}	0.00	126.12	−0.49	−0.01
		1.00×10^{-6}	0.00	−31.60	0.15	0.00
η_2 [h m$^{-1/3}$]	1.00×10^{-5}	1.00×10^{-4}	0.00	173.65	−0.96	−0.56
		1.00×10^{-6}	0.00	−43.71	0.32	0.14
S_s [m^{-1}]	1.60×10^{-4}	1.60×10^{-3}	39.08	0.00	−0.01	0.00
		1.60×10^{-5}	−3.91	0.00	0.00	0.00

RMSE calculated for each of the ParFlow simulations across calibration stages. These are virtually indistinguishable from those obtained with the GPR surrogates. The value of RMSE decreases rapidly to values below 8 m during the first cycle (stages 1–4). In the second cycle (stages 5–8), RMSE stabilizes, suggesting that further calibration cycles are unnecessary.

Conductivity values stemming from calibration align with typical ranges of values for these geomaterial types (Bear et al., 1968). However, the geomaterial class (or macro-category) denoted as gravel appears slightly less conductive than sand, while clay appears to be slightly more conductive than silt. This apparent discrepancy stems from our conceptual model. Specifically, we reduce a complex system comprising about 23 geomaterial classes (as identified in previous studies (Previati et al., 2025)) to six macro-categories. For instance, the macro-category denoted as gravel aggregates all gravel subtypes, including finer components. We further remark that each block of our

Table 3
Parameter values and support ranges (in square brackets) considered during each calibration stage. The number of parameter value instances employed for ParFlow simulations, which form the basis for surrogates, is shown in parentheses after each range.

Stage	k_1 [m h^{-1}]	k_2 [m h^{-1}]	k_3 [m h^{-1}]	k_4 [m h^{-1}]	Optimal value [m h^{-1}]
1	36.0	1.14	1.14×10^{-3}	[0.003–0.114] (6)	k_4 = 9.45×10^{-3}
2	36.0	[0.036–36.0] (7)	1.14×10^{-3}	9.45×10^{-3}	k_2 = 4.46
3	[0.11–114.0] (7)	4.46	1.14×10^{-3}	9.45×10^{-3}	k_1 = 1.79
4	1.79	4.46	[1.14×10^{-4}–1.14×10^{-2}] (5)	9.45×10^{-3}	k_3 = 1.47×10^{-3}
5	1.79	4.46	1.47×10^{-3}	[0.003–0.06] (6)	k_4 = 1.18×10^{-2}
6	1.79	[1.0–8.0] (6)	1.47×10^{-3}	1.18×10^{-2}	k_2 = 4.20
7	[0.5–10.0] (5)	4.20	1.47×10^{-3}	1.18×10^{-2}	k_1 = 1.52
8	1.52	4.20	[3.6×10^{-4}–8.0×10^{-3}] (5)	1.18×10^{-2}	k_3 = 1.26×10^{-3}

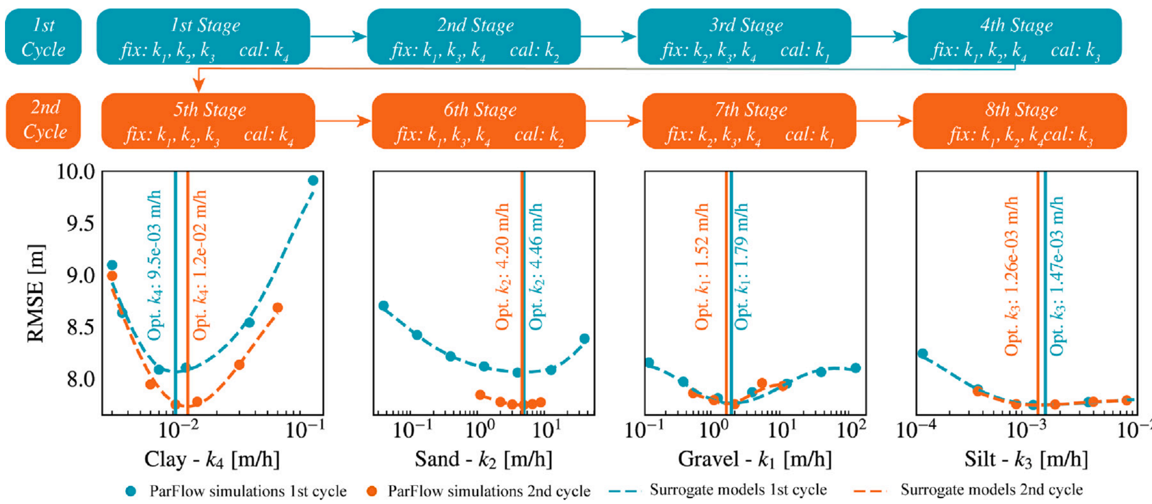

Fig. 5. Stagewise calibration process for conductivity parameters. Each panel depicts RMSE as a function of the facies conductivity under calibration. Symbols correspond to ParFlow-based RMSE values at the sampled points used to train the GPR models; dashed curves denote GPR-based RMSE values. Vertical lines mark the optimal parameter values identified in the first (cyan) and second (orange) calibration cycles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

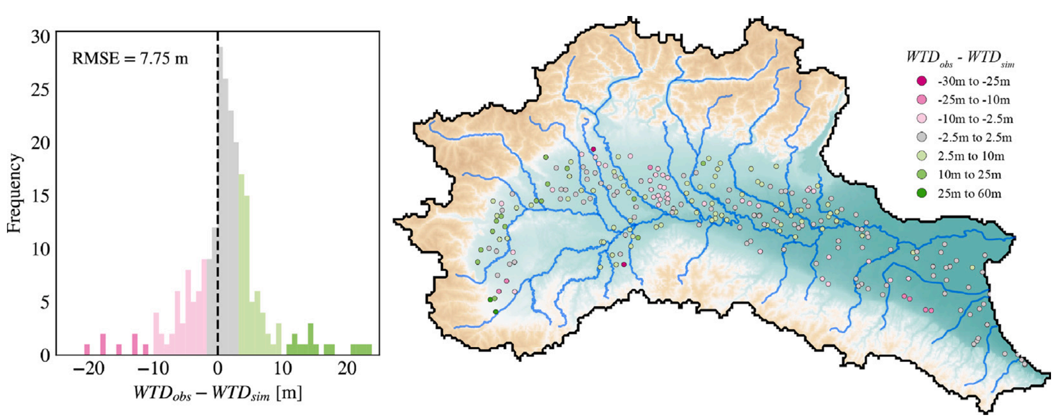

Fig. 6. Histogram (left) and spatial distribution (right) of differences between observed and modeled values of WTD ($WTD_{obs} - WTD_{sim}$) obtained through the calibrated model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

model contains a weighted mixture of all categories, based on their percentage occurrence. Consequently, the calibrated conductivities should be interpreted as effective (or representative) values that reproduce observed WTD, rather than as intrinsic properties of distinct geological units.

Fig. 6 depicts the histogram and spatial distribution of differences (or errors) between observed and modeled values of WTD obtained with the calibrated model. Most errors exceeding 10 m are seen to take place in the foothill area (23 out of 26 points), no clear spatial trend being otherwise observed elsewhere. These large errors are consistent with the higher uncertainty in the geological reconstruction in the mountain area (Manzoni et al., 2023) and with the fractured nature of the system conveying groundwater flow in these areas, which is not reproduced by our modeling approach.

Overall, about 90% of simulated WTD values are characterized by absolute errors below 10 m, 45% of these being less than 2.5 m. We consider this accuracy acceptable given the 2 km horizontal resolution of our model, where each simulated WTD value represents a spatial average of the piezometric surface over a 2 km by 2 km area and therefore cannot match point measurements from individual piezometers. Additional factors contributing to the observed errors likely stem from the fact that our calibration strategy relies on temporally averaged (and

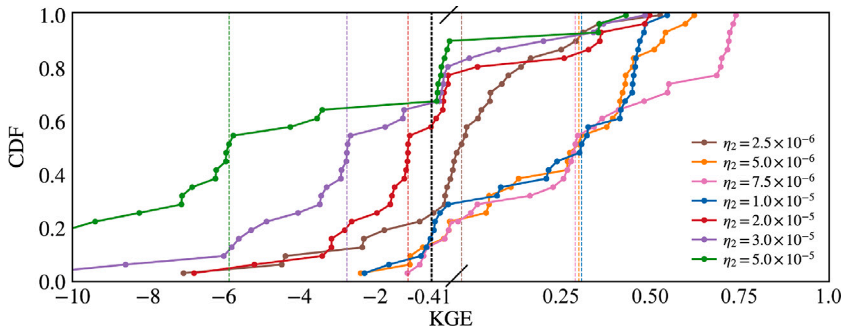

Fig. 7. Cumulative distribution functions (CDFs) of the Kling–Gupta Efficiency (KGE) metric, based on daily river flow rate at 31 gauging stations, for seven values of η_2. Vertical colored dashed lines indicate median KGE values for each η_2 (h m$^{-1/3}$). The black dashed line denotes the threshold KGE = − 0.41. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

otherwise transient) WTD observations for comparison against steady-state model outcomes. The calibrated model also exhibits a slight negative bias (defined as the mean difference between simulated and observed WTD values) of −1.43 m, indicating a slight underestimation of the mean (across the domain) WTD. Negative biases have been reported in previous studies (e.g., Maxwell et al., 2015; Fan et al., 2013) and are typically attributed to overestimated recharge, underestimated lateral flow, or observation wells affected by local pumping.

4.3. Calibration of manning roughness coefficient

In the second phase of the calibration process, we proceed to calibration of the Manning roughness coefficient of the channel cells (η_2). To do so, we sample seven values of η_2 within the interval $[2.5 \times 10^{-6}, 5 \times 10^{-5}]$ h m$^{-1/3}$, spanning conditions from extremely smooth to very rough natural systems (Chow, 1959). We then run ParFlow-CLM simulations under transient conditions at hourly temporal resolution for each of these seven parameter values for the 2009–2018 period. To avoid the influence of initial conditions on calibration results, we conduct a spin-up procedure as described in Section 3.3.3.

Fig. 7 depicts the cumulative distribution functions (CDF) of the Kling–Gupta Efficiency (KGE) metric (Gupta et al., 2009) for the seven values of η_2 considered, based on comparisons between modeled and observed daily river flow rates at the 31 gauging stations reported in Section 3.3.2. To improve readability, different horizontal scales are employed to illustrate negative and positive KGE values, as most of the KGE values fall within the relatively narrow $0 \leq$ KGE < 1 interval. The black vertical dashed line corresponds to KGE = −0.41, the threshold value at which model results describe observations with the same KGE score than the mean observed flow. This is often considered as a threshold for good model performance (Knoben et al., 2019). Fig. 7 also includes vertical colored dashed lines depicting the median KGE value for each η_2. Curves that tend to lie on the right side of the plot are associated with better performing model settings. According to our results, values of η_2 in the range $\eta_2 = [5 \times 10^{-6} - 1 \times 10^{-5}]$ h m$^{-1/3}$ (orange, pink, and blue curves in Fig. 7) tend to feature the highest quality performance. In terms of median KGE, $\eta_2 = 5 \times 10^{-6}$ h m$^{-1/3}$, $\eta_2 = 7.5 \times 10^{-6}$ h m$^{-1/3}$ and $\eta_2 = 1 \times 10^{-5}$ h m$^{-1/3}$ yield nearly identical performance, with a median KGE of 0.35, 0.33, and 0.34, respectively. In terms of percentage of stations for which KGE ≥ -0.41, all of these three values of η_2 produce very similar results, with $\eta_2 = 7.5 \times 10^{-6}$ h m$^{-1/3}$ and $\eta_2 = 5.0 \times 10^{-6}$ h m$^{-1/3}$ being the best performing values (with 87% of the stations being characterized by KGE ≥ -0.41) and $\eta_2 = 1 \times 10^{-5}$ h m$^{-1/3}$ performing slightly less well (with 84% of the stations above the threshold). Notably, $\eta_2 = 7.5 \times 10^{-6}$ h m$^{-1/3}$ yields significantly enhanced performance for some of the stations (associated with KGE values larger than 0.5). Relying on this value of η_2 yields KGE values of 0.75, thus corresponding to a quality level that is not achieved by any of the other Manning roughness coefficient values considered.

The (overall) best performing parameter value (i.e., $\eta_2 = 7.5 \times 10^{-6}$ h m$^{-1/3}$) aligns well with typical Manning roughness coefficients of natural streams (Chow, 1959) and with values documented in the area. For example, Aureli and Mignosa (2004) consider a Manning coefficient of 7.9×10^{-6} h m$^{-1/3}$ to be adequate to describe the roughness of the river and nearby floodplains in a portion of the valley close to the city of Mantua. Nones et al. (2018) adopt a value of 5.5×10^{-6} h m$^{-1/3}$ for a section of the Po River close to the city of Ferrara during a flood simulation. Masoero et al. (2013) estimate a value of 8.3×10^{-6} h m$^{-1/3}$ from a one-dimensional flooding simulation of a 93 km portion of the Po River.

Fig. 8a depicts the spatial distribution of KGE values obtained at each gauging station with $\eta_2 = 7.5 \times 10^{-6}$ h m$^{-1/3}$. Results show a spatial pattern according to which model performance (in terms of KGE) is highest in the valley and Apennine areas, and lowest in the northern and mountain regions. The observed reduced performance in the mountain area is consistent with the results of the first calibration phase. Moreover, the northern areas are heavily impacted by anthropic interventions and artificial reservoirs, factors that are not included in our large-scale model. Specifically, the Po River watershed contains over 2000 reservoirs, mainly concentrated in the north (Molle et al., 2019) with a total storage capacity up to 1000 Mm^3, serving hydropower production, irrigation supply and flooding control.

Fig. 8b complements the analysis by showing the best value of the KGE metric obtained at each of the flow rate stations together with the corresponding value of η_2 (rescaled by a factor of 10^{-5}). These results suggest that the ability of our model to render river flow rates could potentially be improved by adopting a spatially heterogeneous Manning roughness coefficient along the river reaches. All but one of the stations achieve a KGE ≥ 0.2.

5. Conclusions

We introduce an approach for the calibration of large-scale, physically based, fully integrated surface–subsurface (SW–GW) hydrological models. To our knowledge, no other calibrated model of this kind has previously been documented at such spatial scale. Our approach synergistically merges parallel streams of research (i.e., model formulation, sensitivity analysis, model reduction, and calibration). Hence, calibration of a large-scale integrated surface–subsurface hydrological model is made possible using a sequence of complementary methodologies that are here integrated for the first time. These include: (i) a neural network-based probabilistic reconstruct of the internal aquifer architecture that allows reducing geological complexity to a limited set of parameters/facies; (ii) a rigorous sensitivity analysis, leading to identification of model parameters that are most influential on available data associated with quantities of interest (such as, e.g., groundwater levels, discharges in rivers, and/or surface/subsurface water storage);

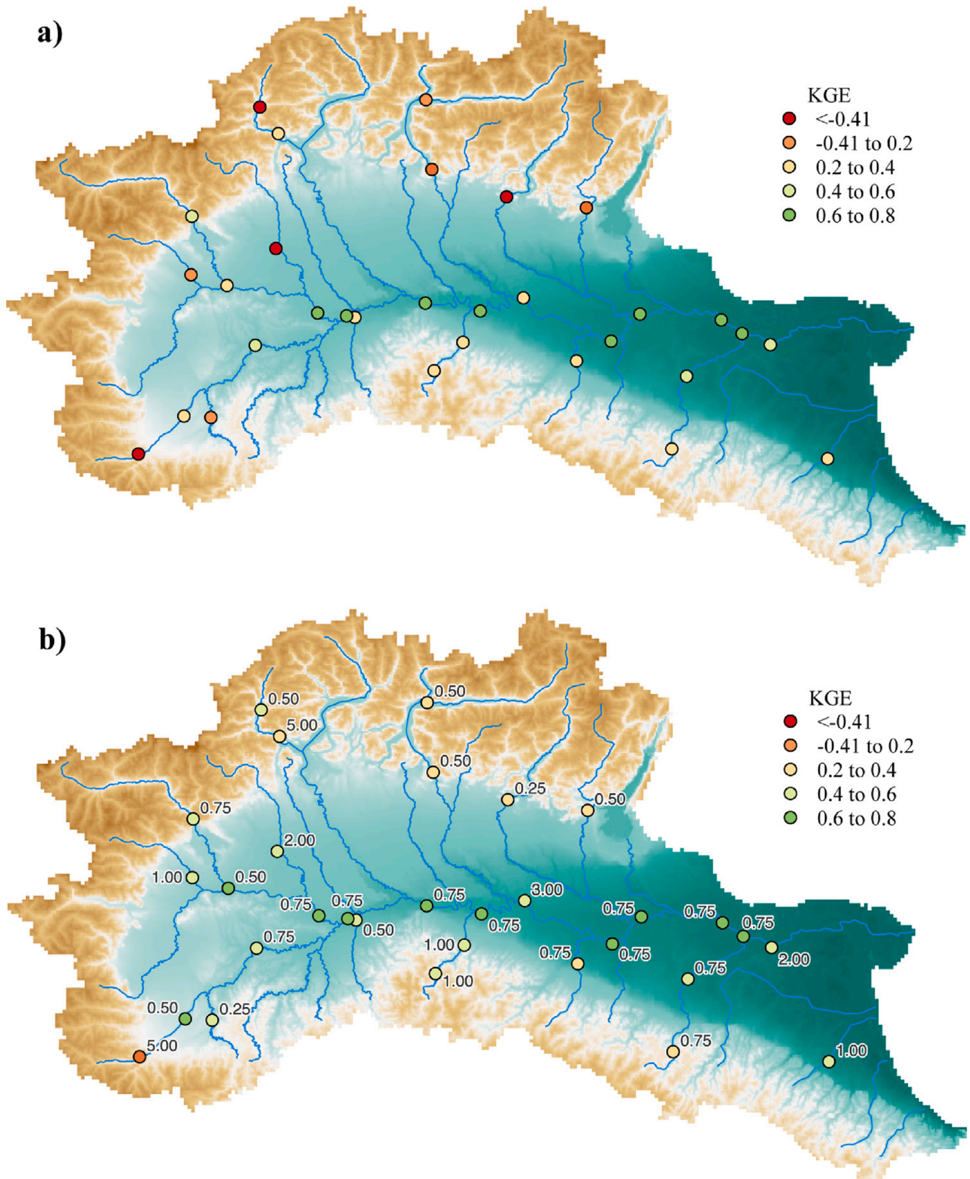

Fig. 8. Spatial distribution of KGE values calculated at each gauging station (a) using a uniform Manning roughness coefficient $\eta_2 = 7.5 \times 10^{-6}$ h m$^{-1/3}$; and (b) optimizing the value of η_2 in each gauging station. Panel (b) also shows the corresponding η_2 values (rescaled by 10^{-5} h m$^{-1/3}$) for each station. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(*iii*) formulation of surrogate models to emulate computationally expensive full-model responses; and (*iv*) a stagewise calibration procedure. Such a unique integration yields efficient parameter estimation in high-dimensional, computationally demanding modeling environments, wherein traditional calibration approaches are infeasible. We demonstrate and document the potential of this framework upon considering the largest basin in Italy, encompassing the Po River District (northern Italy). Our work leads to the following main conclusions:

- A rigorous sensitivity analysis enables one to identify the model parameters that exert the strongest influence on SW–GW dynamics and specific quantities of interest, therefore guiding model calibration efforts. Acquiring such knowledge is critical when assessing fully integrated large-scale models, as simultaneous calibration of all model parameters would be computationally infeasible. In the context of our exemplary application, it enables us to identify the hydraulic conductivity of specific geomaterials/facies and the Manning roughness coefficient of channels as primary controls on subsurface and surface flow behavior, respectively. At the same time, the sensitivity analysis allows highlighting parameters of potential influence that cannot be estimated on the basis of available data, thus guiding future monitoring campaigns. A stark example emerging in our setting is given by porosity and compressibility of certain geomaterials. While they are critical for subsurface storage estimation, they cannot be inferred from standard piezometric campaigns. Recognizing these limitations is particularly important in the context of developing effective water resilience strategy plans, in line with initiatives such as, for example, the 2025 European Water Resilience Strategy (European Commission, 2025).
- Calibration can be performed effectively according to a stepwise process resting on the use of surrogate models. While we rely on Gaussian Process Regression (GPR) surrogates in our setting, other surrogate modeling techniques are fully compatible with the proposed framework. Using GPR, hydraulic conductivities can first be estimated using commonly available steady-state water

table measurements. Model calibration under transient conditions can then be performed in a subsequent stage to refine the estimation of surface roughness coefficients.
- As a direct application, our approach yields the first calibrated, fully integrated, high-fidelity SW–GW model encompassing the entire Po River District. The model offers an unprecedented level of spatial detail and inclusion of physical processes by explicitly representing three-dimensional variably saturated subsurface flow, slope-driven surface runoff, and energy-based estimates of evapotranspiration and snow processes. This comprehensive approach allows for the relaxation of several simplifying assumptions that have commonly been employed in previous large-scale modeling efforts, both in this region and elsewhere.

Future studies include the assessment of the way model calibration could be enhanced using additional available datasets. These may include, for example, satellite-derived evapotranspiration, soil moisture estimates, snow water equivalent maps, and/or terrestrial water storage anomalies (Sun et al., 2016). Integration of these datasets within a multi-stage calibration framework based on the use of multiple datasets could improve model parameter assessment and refine process representation across hydrological compartments. Beyond enhancing predictive accuracy, such efforts may also reveal critical feedbacks between surface and subsurface flow systems, thus further improving our knowledge of large-scale water dynamics. Additional future research avenues include the possibility of informing sensitivity analyses through parameter values (and their uncertainty bounds) stemming from model calibration. In this context, setting (stochastic) model calibration in frameworks relying, for example, on Markov-Chain Monte Carlo approaches can benefit from the availability of surrogate models to reduce computational burdens. Our results further suggest that additional research efforts could focus on investigating, for example, the effects of considering reach-scale-heterogeneous Manning roughness coefficients (i.e., values of the Manning coefficients varying from one river reach to another while remaining uniform within each reach). Such an approach may improve the representation of spatial variability in channel characteristics and, consequently, yield an overall enhanced model performance.

CRediT authorship contribution statement

Leonardo Sandoval: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation. **Alberto Guadagnini:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis, Conceptualization. **Laura Condon:** Writing – review & editing, Supervision, Methodology. **Monica Riva:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Financial support of the European Union’s Horizon 2020 research and innovation program through the RECYCLE project under the Marie Skłodowska-Curie Grant Agreement 872607 is acknowledged. M. Riva acknowledges support from Water Alliance - Acque di Lombardia. Funding from the European Union Next Generation EU (National Recovery and Resilience Plan - NRRP, Mission 4, Component 2, Investment 1.3 - D.D. 1243 2/8/2022, PE0000005) in the context of the RETURN Extended Partnership is acknowledge. We acknowledge ISCRA for awarding this project access to GALILEO100 and LEONARDO supercomputers, owned by the EuroHPC Joint Undertaking and hosted by CINECA (Italy).

Data availability

Data will be made available on request.

References

Agrawala, S., 2007. Climate Change in the European Alps - Adapting Winter Tourism and Natural Hazards Management. Organisation for Economic Cooperation and Development, http://dx.doi.org/10.1787/9789264031692-en.

Anderson, M.P., Woessner, W.W., Hunt, R.J., 2015. Applied Groundwater Modeling: Simulation of Flow and Advective Transport. Academic Press, http://dx.doi.org/10.1016/C2009-0-21563-7.

Antonellini, M., Mollema, P., Giambastiani, B., Bishop, K., Caruso, L., Minchio, A., Pellegrini, L., Sabia, M., Ulazzi, E., Gabbianelli, G., et al., 2008. Salt water intrusion in the coastal aquifer of the southern Po Plain, Italy. Hydrogeol. J. 16 (8), 1541–1556. http://dx.doi.org/10.1007/s10040-008-0319-9.

Aureli, F., Mignosa, P., 2004. Flooding scenarios due to levee breaking in the Po river. In: Proceedings of the Institution of Civil Engineers - Water Management. Vol. 157, pp. 3–12. http://dx.doi.org/10.1680/wama.2004.157.1.3.

Autorita di Bacino del Fiume Po, 2012. Piano di Bilancio Idrico del Distretto Idrografico del Fiume Po - Valutazione Ambientale Stategica. Technical Report, Autorita di Bacino del Fiume Po, URL: va.mite.gov.it/it-IT/Oggetti/Documentazione/1451/2200.

Autorita di Bacino del Fiume Po, 2021. Piano di Gestione Del Distretto Idrografico del Fiume Po Al 2021. Technical Report, Autorita di Bacino del Fiume Po, URL: www.adbpo.it/PianoAcque2021.

Batu, V., 1998. Aquifer Hydraulics: a Comprehensive Guide to Hydrogeologic Data Analysis. John Wiley & Sons.

Bear, J., 2007. Hydraulics of Groundwater. Dover Publications, Inc..

Bear, J., Cheng, A.H.-D., 2010. Modeling Groundwater Flow and Contaminant Transport. Springer, http://dx.doi.org/10.1007/978-1-4020-6682-5.

Bear, J., Zaslavsky, D., Irmay, S., 1968. Physical Principles of Water Percolation and Seepage. United Nations Educational, Scientific and Cultural Organisation, http://dx.doi.org/10.2136/sssaj1968.03615995003200060003x.

Beniston, M., Stephenson, D.B., Christensen, O.B., Ferro, C.A., Frei, C., Goyette, S., Halsnaes, K., Holt, T., Jylhä, K., Koffi, B., et al., 2007. Future extreme events in European climate: an exploration of regional climate model projections. Clim. Change 81, 71–95. http://dx.doi.org/10.1007/s10584-006-9226-z.

Bollmeyer, C., Keller, J., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., et al., 2015. Towards a high-resolution regional reanalysis for the European CORDEX domain. Q. J. R. Meteorol. Soc. 141 (686), 1–15. http://dx.doi.org/10.1002/qj.2486.

Bozzola, M., Swanson, T., 2014. Policy implications of climate variability on agriculture: Water management in the Po River basin, Italy. Environ. Sci. Policy 43, 26–38. http://dx.doi.org/10.1016/j.envsci.2013.12.002.

Carlson, G., Massari, C., Rotiroti, M., Bonomi, T., Preziosi, E., Wilder, A., Whitaker, D., Girotto, M., 2025. Intensive irrigation buffers groundwater declines in key European breadbasket. Nat. Water 1–10. http://dx.doi.org/10.1038/s44221-025-00445-4.

Carrera, J., Neuman, S.P., 1986. Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. Water Resour. Res. 22 (2), 199–210. http://dx.doi.org/10.1029/WR022i002p00199.

Center For International Earth Science Information Network-CIESIN-Columbia University, 2018. Gridded population of the world (GPWv4), land and water area, revision 11 [data set]. http://dx.doi.org/10.7927/H49C6VHW.

Chow, V.T., 1959. Open-Channel Hydraulics. McGraw-Hill Book Company.

Christensen, J.H., Christensen, O.B., 2003. Severe summertime flooding in Europe. Nature 421 (6925), 805–806. http://dx.doi.org/10.1038/421805a.

Cislaghi, A., Masseroni, D., Massari, C., Camici, S., Brocca, L., 2020. Combining a rainfall–runoff model and a regionalization approach for flood and water resource assessment in the western Po Valley, Italy. Hydrol. Sci. J. 65 (3), 348–370. http://dx.doi.org/10.1080/02626667.2019.1690656.

Colombani, N., Osti, A., Volta, G., Mastrocicco, M., 2016. Impact of climate change on salinization of coastal water resources. Water Resour. Manag. 30 (7), 2483–2496. http://dx.doi.org/10.1007/s11269-016-1292-z.

Condon, L.E., Kollet, S., Bierkens, M.F., Fogg, G.E., Maxwell, R.M., Hill, M.C., Fransen, H.-J.H., Verhoef, A., Van Loon, A.F., Sulis, M., et al., 2021. Global groundwater modeling and monitoring: Opportunities and challenges. Water Resour. Res. 57 (12), e2020WR029500. http://dx.doi.org/10.1029/2020WR029500.

Condon, L.E., Maxwell, R.M., 2019. Modified priority flood and global slope enforcement algorithm for topographic processing in physically based hydrologic modeling applications. Comput. Geosci. 126, 73–83. http://dx.doi.org/10.1016/j.cageo.2019.01.020.

Coppola, E., Verdecchia, M., Giorgi, F., Colaiuda, V., Tomassetti, B., Lombardi, A., 2014. Changing hydrological conditions in the Po basin under global warming. Sci. Total Environ. 493, 1183–1196. http://dx.doi.org/10.1016/j.scitotenv.2014.03.003.

Dai, Y., Zeng, X., Dickinson, R., Baker, I., Bonan, G., Bosilovich, M., Denning, A., Dirmeyer, P., Houser, P., Niu, G., et al., 2003. The common land model. Bull. Am. Meteorol. Soc. 84, 1013–1024. http://dx.doi.org/10.1175/BAMS-84-8-1013.

Dall'Amico, M., Tasin, S., Di Paolo, F., Brian, M., Leoni, P., Tornatore, F., Formetta, G., Wani, J.M., Rigon, R., Roati, G., 2025. 30-Years (1991–2021) snow water equivalent dataset in the Po River District, Italy. Sci. Data 12 (1), 374. http://dx.doi.org/10.1038/s41597-025-04633-5.

Dari, J., Brocca, L., Modanesi, S., Massari, C., Tarpanelli, A., Barbetta, S., Quast, R., Vreugdenhil, M., Freeman, V., Barella-Ortiz, A., et al., 2023. Regional data sets of high-resolution (1 and 6 km) irrigation estimates from space. Earth Syst. Sci. Data 15 (4), 1555–1575. http://dx.doi.org/10.5194/essd-15-1555-2023.

de Graaf, I.E., van Beek, R.L., Gleeson, T., Moosdorf, N., Schmitz, O., Sutanudjaja, E.H., Bierkens, M.F., 2017. A global-scale two-layer transient groundwater model: Development and application to groundwater depletion. Adv. Water Resour. 102, 53–67. http://dx.doi.org/10.1016/j.advwatres.2017.01.011.

Elsasser, H., Bürki, R., 2002. Climate change as a threat to tourism in the Alps. Clim. Res. 20 (3), 253–257. http://dx.doi.org/10.3354/cr020253.

Engdahl, N.B., 2024. Impacts of permeability uncertainty in a coupled surface-subsurface flow model under perturbed recharge scenarios. Water Resour. Res. 60 (3), e2023WR035975. http://dx.doi.org/10.1029/2023WR035975.

European Commission, 2025. Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions on a European water resilience strategy. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52025DC0280.

Famiglietti, J.S., 2014. The global groundwater crisis. Nat. Clim. Chang. 4 (11), 945–948. http://dx.doi.org/10.1038/nclimate2425.

Fan, Y., Li, H., Miguez-Macho, G., 2013. Global patterns of groundwater table depth. Science 339 (6122), 940–943. http://dx.doi.org/10.1126/science.1229881.

Foster, L., Maxwell, R., 2019. Sensitivity analysis of hydraulic conductivity and Manning's n parameters lead to new method to scale effective hydraulic conductivity across model resolutions. Hydrol. Process. 33 (3), 332–349. http://dx.doi.org/10.1002/hyp.13327.

Fratianni, S., Acquaotta, F., 2017. The Climate of Italy. Springer, pp. 29–38. http://dx.doi.org/10.1007/978-3-319-26194-2_4.

Friedl, M., Sulla-Menashe, D., 2019. MODIS/Terra+Aqua land cover type yearly L3 global 500m SIN grid V006 [data set]. http://dx.doi.org/10.5067/MODIS/MCD12Q1.006.

Gleeson, T., VanderSteen, J., Sophocleous, M.A., Taniguchi, M., Alley, W.M., Allen, D.M., Zhou, Y., 2010. Groundwater sustainability strategies. Nat. Geosci. http://dx.doi.org/10.1038/ngeo881.

Gleeson, T., Wagener, T., Döll, P., Zipper, S.C., West, C., Wada, Y., Taylor, R., Scanlon, B., Rosolem, R., Rahman, S., et al., 2021. GMD perspective: The quest to improve the evaluation of groundwater representation in continental-to global-scale models. Geosci. Model. Dev. 14 (12), 7545–7571. http://dx.doi.org/10.5194/gmd-14-7545-2021.

Grimm, M., Jones, R.J., Rusco, E., Montanarella, L., 2003. Soil Erosion Risk in Italy: A Revised USLE Approach. Technical Report, European Soil Bureau Research, URL: https://esdac.jrc.ec.europa.eu/ESDB_Archive/pesera/pesera_cd/pdf/ita_er06.pdf.

Guadagnini, L., Menafoglio, A., Sanchez-Vila, X., Guadagnini, A., 2020. Probabilistic assessment of spatial heterogeneity of natural background concentrations in large-scale groundwater bodies through functional geostatistics. Sci. Total Environ. 740, 140139. http://dx.doi.org/10.1016/j.scitotenv.2020.140139.

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. J. Hydrol. 377 (1–2), 80–91. http://dx.doi.org/10.1016/j.jhydrol.2009.08.003.

Istituto Nazionale di Statistica, 2020. Public water supply use [data set]. URL: http://dati.istat.it/.

Janetti, E.B., Riva, M., Guadagnini, A., 2021. Natural springs protection and probabilistic risk assessment under uncertain conditions. Sci. Total Environ. 751, 141430. http://dx.doi.org/10.1016/j.scitotenv.2020.141430.

Jefferson, J.L., Maxwell, R.M., Constantine, P.G., 2017. Exploring the sensitivity of photosynthesis and stomatal resistance parameters in a land surface model. J. Hydrometeorol. 18 (3), 897–915. http://dx.doi.org/10.1175/JHM-D-16-0053.1.

Knoben, W.J., Freer, J.E., Woods, R.A., 2019. Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrol. Earth Syst. Sci. 23 (10), 4323–4331. http://dx.doi.org/10.5194/hess-23-4323-2019.

Kollet, S.J., Maxwell, R.M., 2006. Integrated surface–groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model. Adv. Water Resour. 29 (7), 945–958. http://dx.doi.org/10.1016/j.advwatres.2005.08.006.

Kollet, S.J., Maxwell, R.M., 2008. Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model. Water Resour. Res. 44 (2), http://dx.doi.org/10.1029/2007WR006004.

Lehner, B., Roth, A., Huber, M., Anand, M., Grill, G., Osterkamp, N., Tubbesing, R., Warmedinger, L., Thieme, M., 2021. HydroSHEDS v2. 0-refined global river network and catchment delineations from TanDEM-X elevation data. In: EGU General Assembly Conference Abstracts. pp. EGU21–9277. http://dx.doi.org/10.5194/egusphere-egu21-9277.

Liu, Z., Todini, E., 2002. Towards a comprehensive physically-based rainfall-runoff model. Hydrol. Earth Syst. Sci. 6 (5), 859–881. http://dx.doi.org/10.5194/hess-6-859-2002.

Luo, J., Straffelini, E., Bozzolan, M., Zheng, Z., Tarolli, P., 2024. Saltwater intrusion in the Po River Delta (Italy) during drought conditions: Analyzing its spatio-temporal evolution and potential impact on agriculture. Int. Soil Water Conserv. Res. 12 (3), 714–725. http://dx.doi.org/10.1016/j.iswcr.2023.09.009.

Manzoni, A., Porta, G.M., Guadagnini, L., Guadagnini, A., Riva, M., 2023. Probabilistic reconstruction via machine-learning of the po watershed aquifer system (Italy). Hydrogeol. J. 31 (6), 1547–1563. http://dx.doi.org/10.1007/s10040-023-02677-8.

Manzoni, A., Porta, G.M., Guadagnini, L., Guadagnini, A., Riva, M., 2024. A comprehensive framework for stochastic calibration and sensitivity analysis of large-scale groundwater models. Hydrol. Earth Syst. Sci. 28 (12), 2661–2682. http://dx.doi.org/10.5194/hess-28-2661-2024.

Masoero, A., Claps, P., Asselman, N.E., Mosselman, E., Di Baldassarre, G., 2013. Reconstruction and analysis of the Po River inundation of 1951. Hydrol. Process. 27 (9), 1341–1348. http://dx.doi.org/10.1002/hyp.9558.

Maxwell, R.M., 2013. A terrain-following grid transform and preconditioner for parallel, large-scale, integrated hydrologic modeling. Adv. Water Resour. 53, 109–117. http://dx.doi.org/10.1016/j.advwatres.2012.10.001.

Maxwell, R.M., Condon, L.E., 2016. Connections between groundwater flow and transpiration partitioning. Science 353 (6297), 377–380. http://dx.doi.org/10.1126/science.aaf7891.

Maxwell, R., Condon, L., Kollet, S., 2015. A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3. Geosci. Model. Dev. 8 (3), 923–937. http://dx.doi.org/10.5194/gmd-8-923-2015.

Maxwell, R.M., Kollet, S.J., Condon, L.E., Smith, S.G., Woodward, C.S., Falgout, R.D., Ferguson, I.M., Engdahl, N., Hokkanen, J., Artavanis, G., West, B., Yang, C., Hector, B., Gilbert, J., Bearup, L., Jefferson, J., Baldwin, C., Bosl, W.J., Hornung, R., Ashby, S., Kulkarni, K.B., 2024. Parflow user's manual. URL: https://parflow.readthedocs.io.

Molle, F., Sanchis-Ibor, C., Avellà-Reus, L., 2019. Irrigation in the Mediterranean. Springer, http://dx.doi.org/10.1007/978-3-030-03698-0.

Montanari, A., Nguyen, H., Rubinetti, S., Ceola, S., Galelli, S., Rubino, A., Zanchettin, D., 2023. Why the 2022 Po River drought is the worst in the past two centuries. Sci. Adv. 9 (32), eadg8304. http://dx.doi.org/10.1126/sciadv.adg8304.

Morgan, G.M., 1973. A general description of the hail problem in the Po Valley of northern Italy. J. Appl. Meteorol. Clim. 12 (2), 338–353. http://dx.doi.org/10.1175/1520-0450(1973)012<0338:AGDOTH>2.0.CO;2.

Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., Niemann, C., Peiris, T.A., Popat, E., Portmann, F.T., Reinecke, R., et al., 2021. The global water resources and use model WaterGAP v2. 2d: Model description and evaluation. Geosci. Model. Dev. 14 (2), 1037–1079. http://dx.doi.org/10.5194/gmd-14-1037-2021.

do Nascimento, T.V., Rudlang, J., Höge, M., van der Ent, R., Chappon, M., Seibert, J., Hrachowitz, M., Fenicia, F., 2024. Estreams: An integrated dataset and catalogue of streamflow, hydro-climatic and landscape variables for Europe. Sci. Data 11 (1), 879. http://dx.doi.org/10.1038/s41597-024-03706-1.

Naz, B.S., Sharples, W., Ma, Y., Goergen, K., Kollet, S., 2023. Continental-scale evaluation of a fully distributed coupled land surface and groundwater model, ParFlow-CLM (v3. 6.0), over Europe. Geosci. Model. Dev. 16 (6), 1617–1639. http://dx.doi.org/10.5194/gmd-16-1617-2023.

Nones, M., Pugliese, A., Domeneghetti, A., Guerrero, M., 2018. Po river morphodynamics modelled with the open-source code iRIC. In: Free Surface Flows and Transport Processes: 36th International School of Hydraulics. Springer, pp. 335–346. http://dx.doi.org/10.1007/978-3-319-70914-7_22.

Pettine, M., Bianchi, M., Martinotti, W., Muntau, H., Renoldi, M., Tartari, G., 1996. Contribution of the Lambro river to the total pollutant transport in the Po watershed (Italy). Sci. Total Environ. 192 (3), 275–297. http://dx.doi.org/10.1016/S0048-9697(96)05332-6.

Pieri, L., Ventura, F., Ventura, M., Tagliavini, M., Ponti, M., Pistocchi, R., Albertazzi, S., Vignudelli, M., Pisa, P.R., 2011. Chemical and biological indicators of water quality in three agricultural watersheds of the Po valley, Italy. Ital. J. Agron. 6 (4), e36. http://dx.doi.org/10.4081/ija.2011.e36.

Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G.B., Kempen, B., Ribeiro, E., Rossiter, D., 2021. Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240. http://dx.doi.org/10.5194/soil-7-217-2021.

Previati, A., Silvestri, V., Crosta, G., 2025. Deep learning text classification of borehole logs for regional scale modeling of hydrofacies (Po Plain, N Italy). J. Hydrol.: Reg. Stud. 58, 102157. http://dx.doi.org/10.1016/j.ejrh.2024.102157.

Ravazzani, G., Barbero, S., Salandin, A., Senatore, A., Mancini, M., 2015. An integrated hydrological model for assessing climate change impacts on water resources of the upper Po River basin. Water Resour. Manag. 29, 1193–1215. http://dx.doi.org/10.1007/s11269-014-0868-8.

Reinecke, R., Foglia, L., Mehl, S., Herman, J.D., Wachholz, A., Trautmann, T., Döll, P., 2019. Spatially distributed sensitivity of simulated global groundwater heads and flows to hydraulic conductivity, groundwater recharge, and surface water body parameterization. Hydrol. Earth Syst. Sci. 23 (11), 4561–4582. http://dx.doi.org/10.5194/hess-23-4561-2019.

Reinecke, R., Gnann, S., Stein, L., Bierkens, M., de Graaf, I., Gleeson, T., Oude Essink, G., Sutanudjaja, E.H., Vargas, C.R., Verkaik, J., et al., 2024. Uncertainty in model estimates of global groundwater depth. Environ. Res. Lett. 19 (11), 114066. http://dx.doi.org/10.1088/1748-9326/ad8587.

Ryken, A., Bearup, L.A., Jefferson, J.L., Constantine, P., Maxwell, R.M., 2020. Sensitivity and model reduction of simulated snow processes: Contrasting observational and parameter uncertainty to improve prediction. Adv. Water Resour. 135, 103473. http://dx.doi.org/10.1016/j.advwatres.2019.103473.

Saltelli, A., 1999. Sensitivity analysis: Could better methods be used? J. Geophys. Res.: Atmospheres 104 (D3), 3789–3793. http://dx.doi.org/10.1029/1998JD100042.

Sanchez-Vila, X., Guadagnini, A., Carrera, J., 2006. Representative hydraulic conductivities in saturated groundwater flow. Rev. Geophys. 44 (3), http://dx.doi.org/10.1029/2005RG000169.

Scanlon, B.R., Fakhreddine, S., Rateb, A., de Graaf, I., Famiglietti, J., Gleeson, T., Grafton, R.Q., Jobbagy, E., Kebede, S., Kolusu, S.R., et al., 2023. Global water resources and the role of groundwater in a resilient water future. Nat. Rev. Earth Environ. 4 (2), 87–101. http://dx.doi.org/10.1038/s43017-022-00378-6.

Schulzweida, U., 2023. CDO user guide. http://dx.doi.org/10.5281/zenodo.10020800.

Seck, A., Welty, C., Maxwell, R.M., 2015. Spin-up behavior and effects of initial conditions for an integrated hydrologic model. Water Resour. Res. 51 (4), 2188–2210. http://dx.doi.org/10.1002/2014WR016371.

Siena, M., Riva, M., 2020. Impact of geostatistical reconstruction approaches on model calibration for flow in highly heterogeneous aquifers. Stoch. Environ. Res. Risk Assess. 34 (10), 1591–1606. http://dx.doi.org/10.1007/s00477-020-01865-2.

Spinoni, J., Vogt, J.V., Naumann, G., Barbosa, P., Dosio, A., 2018. Will drought events become more frequent and severe in Europe? Int. J. Climatol. 38 (4), 1718–1736. http://dx.doi.org/10.1002/joc.5291.

Stramondo, S., Saroli, M., Tolomei, C., Moro, M., Doumaz, F., Pesci, A., Loddo, F., Baldi, P., Boschi, E., 2007. Surface movements in bologna (po plain—Italy) detected by multitemporal DInSAR. Remote Sens. Environ. 110 (3), 304–316. http://dx.doi.org/10.1016/j.rse.2007.02.023.

Sun, Y., Riva, R., Ditmar, P., 2016. Optimizing estimates of annual variations and trends in geocenter motion and J2 from a combination of GRACE data and geophysical models. J. Geophys. Res.: Solid Earth 121 (11), 8352–8370. http://dx.doi.org/10.1002/2016JB013073.

Terzaghi, K., Peck, R.B., Mesri, G., 1996. Soil Mechanics in Engineering Practice. John Wiley & Sons.

Van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Am. J. 44 (5), 892–898. http://dx.doi.org/10.2136/sssaj1980.03615995004400050002x.

Verkaik, J., Sutanudjaja, E.H., Oude Essink, G.H., Lin, H.X., Bierkens, M.F., 2024. GLOBGM v1. 0: a parallel implementation of a 30 arcsec PCR-GLOBWB-MODFLOW global-scale groundwater model. Geosci. Model. Dev. 17 (1), 275–300. http://dx.doi.org/10.5194/gmd-17-275-2024.

Vezzoli, R., Mercogliano, P., Pecora, S., Zollo, A., Cacciamani, C., 2015. Hydrological simulation of Po River (North Italy) discharge under climate change scenarios using the RCM COSMO-CLM. Sci. Total Environ. 521, 346–358. http://dx.doi.org/10.1016/j.scitotenv.2015.03.096.

Viaroli, P., Soana, E., Pecora, S., Laini, A., Naldi, M., Fano, E.A., Nizzoli, D., 2018. Space and time variations of watershed N and P budgets and their relationships with reactive N and P loadings in a heavily impacted river basin (Po River, Northern Italy). Sci. Total Environ. 639, 1574–1587. http://dx.doi.org/10.1016/j.scitotenv.2018.05.233.

Xanke, J., Liesch, T., 2022. Quantification and possible causes of declining groundwater resources in the Euro-Mediterranean region from 2003 to 2020. Hydrogeol. J. 30 (2), 379–400. http://dx.doi.org/10.1007/s10040-021-02448-3.

Xie, Z., Yao, Y., Zhang, X., Liang, S., Fisher, J.B., Chen, J., Jia, K., Shang, K., Yang, J., Yu, R., et al., 2022. The global land surface satellite (glass) evapotranspiration product version 5.0: Algorithm development and preliminary validation. J. Hydrol. 610, 127990. http://dx.doi.org/10.1016/j.jhydrol.2022.127990.

Yang, C., Tijerina-Kreuzer, D.T., Tran, H.V., Condon, L.E., Maxwell, R.M., 2023. A high-resolution, 3D groundwater-surface water simulation of the contiguous US: Advances in the integrated ParFlow CONUS 2.0 modeling platform. J. Hydrol. 626, 130294. http://dx.doi.org/10.1016/j.jhydrol.2023.130294.