

To tame a land: Limiting factors in model performance for the multi-objective calibration of a pan-European, semi-distributed hydrological model for discharge and sediments

Conrad Brendel^{*}, René Capell, Alena Bartosova

Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

ARTICLE INFO

Keywords:

Calibration
Large-Scale
HYPE
Hydrological Model
Europe
Sediment

ABSTRACT

Study Region: Europe

Study Focus: The semi-distributed Hydrological Predictions for the Environment (HYPE) model for the European domain, E-HYPE4, was calibrated for discharge and sediments within a framework allowing for evaluation of the factors limiting model performance. Calibration was conducted using a multi-phase approach during which an initial multi-objective discharge/sediment calibration was followed by an exhaustive sediment calibration in which combinations of sediment routines were assessed. During each calibration phase, ensembles of parameter sets and model routines were simultaneously evaluated against discharge, evapotranspiration, and sediment observations. In total, 20,000 parameter sets were evaluated during the discharge calibration, and a further 20,000 model setups were assessed during the sediment calibration.

New Hydrological Insights for the Region: Model performance was best with a highly regionalized model, and the largest drop in achievable performance occurred when transitioning from an ensemble of candidates to a single model setup. Much of the performance gains from a highly regionalized model could, however, be achieved with a much less regionalized model. Inclusion of sediments as a calibration objective provided more value than that of evapotranspiration in regards to reducing equifinality in the calibration for discharge. Evaluation of the erosion routines indicated that an index-based routine performed equally well as a more complex, process-based routine. Finally, analysis of model performance by subbasin attributes revealed the dominant factors — such as landuse, glaciers, abstractions/regulations, groundwater, and lakes/wetlands — affecting model biases for various regions.

1. Introduction

The ultimate goal in hydrological modeling is to tame a land — that is, to capture the processes and dynamics of hydrological systems with such accuracy that we allow ourselves the power to foresee or to look into the past. Through operational forecasting, conducting climate impact assessments, generating inputs to oceanographic models, analyzing historical events and periods, and more, large-scale hydrological models provide information to meet the needs of society and the environment. However, decisions made regarding model structure, parameterization, and evaluation can impose limits to this power in terms of model performance,

Abbreviations: SLC, Soil type - Landuse Combination; MMF, Morgan-Morgan-Finney; RGB, Representative Gauged Basin.

^{*} Corresponding author.

E-mail address: conrad.brendel@smhi.se (C. Brendel).

<https://doi.org/10.1016/j.ejrh.2023.101544>

Received 29 May 2023; Received in revised form 4 October 2023; Accepted 5 October 2023

Available online 11 October 2023

2214-5818/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

uncertainty, and applicability. Thus, to responsibly wield this power, we must understand the impacts of these decisions as well as the implications of these limitations for subsequent model applications.

During model development, modelers must make many important choices and assumptions in regards to model structure. Model structure is multifaceted and includes the choice of modeled processes, the process descriptions and representations, how the processes are coupled, the numerical discretization of the equations, the model spatial representation, and the interpretation and classification of geographic features such as soil type, landuse, crops, etc. (Butts et al., 2004). In large-scale modeling, these choices and assumptions may often be constrained by data availability and/or computation requirements. Many studies have shown that model performance does not necessarily improve with increased structural and parametric complexity (Brendel et al., 2021; Finger et al., 2015; Orth et al., 2015; Troin et al., 2022; van Esse et al., 2013). However, the performance of model structures can vary in certain regions and hydrological conditions, and differences in model structures can result in significant differences in simulated hydrological responses, especially for extreme events (Bartosova et al., 2021; Karlsson et al., 2016; Troin et al., 2022; van Kempen et al., 2021). Thus, the explicit and implicit decisions made regarding model structure have important consequences for ensuing model usage.

When parameterizing large-scale hydrological models, modelers must often choose between two opposing paradigms: a global approach or a regionalized approach. In a global approach, one set of model parameters is selected and applied to describe all processes for all spatial units (e.g. catchments) in the model domain based on pre-selected characteristics such as land use or soil type. In contrast, with a regionalized approach, model parameters are estimated for each spatial unit in a carefully selected calibration domain and applied to the remainder of the model domain according to the similarity with the areas where the model was calibrated. Strategies to transfer parameters from the calibration domain include those based on spatial proximity, physical similarity of climate and catchment characteristics, weighted averaging of parameter sets, multiple linear regression of parameters and catchment descriptors, and regional calibration (Arsenault et al., 2019; Blöschl, 2013; Mizukami et al., 2017; Parajka et al., 2013; Qi et al., 2021). Whereas the methods based on spatial proximity, similarity, and weighted averaging apply the entire parameter set from a gauged basin to an ungauged basin, the methods based on parameter regression and regional calibration relate individual model parameters to the characteristics of the domain (Parajka et al., 2013). The choice of regionalization method can significantly impact the uncertainty of the model results (Qi et al., 2021), and this can have important consequences for subsequent modeling activities such as climate impact projections (Ashraf Vaghefi et al., 2019).

Equifinality occurs when various distinct model configurations — model structures, parameter sets, performance metrics, initial/boundary conditions, inputs, and internal fluxes — provide equally acceptable performance (Khatami et al., 2019). Managing equifinality is an inevitable challenge faced by modelers during calibration of large-scale hydrological models. A common strategy for reducing equifinality is to introduce additional target variables to the calibration in order to improve the representation of specific processes in the hydrologic cycle. Earth observations can provide such additional calibration data for large spatial domains, and multi-objective calibrations have been performed with discharge and data linked to vegetation (Fernandez-Palomino et al., 2021; Rane and Jayaraj, 2022), evapotranspiration (Liu et al., 2022; Rane and Jayaraj, 2022; Shah et al., 2021; Sirisena et al., 2020), snow (Di Marco et al., 2021; Finger et al., 2015), and soil moisture (Eini et al., 2023; Yang et al., 2019). Model calibrations have also been constrained using in-situ observations for sediments (de Lavenne et al., 2022), soil moisture (Budhathoki et al., 2020), snow water equivalent (Nemri and Kinnard, 2020), and nutrients (Ghaffar et al., 2021). Another strategy for managing equifinality is to include additional calibration metrics, such as those for various parts of the flow regime or the timing of the hydrograph (Casado-Rodríguez and del Jesus, 2022; Fernandez-Palomino et al., 2021; Her and Seong, 2018; Zhang et al., 2016). Some level of equifinality is unavoidable during model calibration, and the resulting uncertainty in the model predictions can raise questions regarding the reliability of the model for forecasting (Yatheendradas et al., 2008) and applicability of the model for climate impact analyses (Her et al., 2019; Krysanova et al., 2018; Wilby, 2005).

In this paper, we present insights gained on the factors limiting the performance of large-scale hydrological models during the multi-objective calibration of a pan-European, semi-distributed hydrological model for discharge and sediments. The Hydrological Predictions for the Environment (HYPE) model setup for Europe, referred to as E-HYPE, was developed to perform operational forecasting, estimate discharges and loads for oceanographic models, conduct climate change impact assessments, and assist with assessment of mitigation scenarios and other scientific research. The second- and third-generation versions of the E-HYPE model setup are described in Donnelly et al. (2016) and Hundecha et al. (2016), respectively. Here, we describe the development and calibration of the new, fourth-generation version of E-HYPE, E-HYPE4, which provides improved catchment delineations, new landuse and meteorological forcing data, and adds sediment simulations. Overall, the aim of this study was to create a flexible, traceable calibration strategy with which we could assess the impacts of data availability, model structure, process descriptions, target calibration variables, and objective criteria on equifinality and model performance. Specifically, our objectives were to:

- Investigate the impact of data availability on the transferability of model parameters from gauged to ungauged catchments
- Identify the dominant processes limiting model performance for various geographic regions and prevailing catchment characteristics
- Evaluate the significance of the selection and process representation of available erosion model and in-stream sedimentation/resuspension routines for model performance
- Assess the limits of achievable model performance with varying degrees of model regionalization
- Compare the value of incorporating evapotranspiration and sediment data in reducing equifinality in multi-objective calibrations

2. Data & Methods

2.1. Hydrological Model

The HYPE model is an open source, physically-based, semi-distributed model developed by the Swedish Meteorological and Hydrological Institute to simulate water quantity and quality (SMHI, 2023). HYPE's spatial representation consists of subbasins characterized according to hydrological response units called Soil type – Landuse Combination (SLC) classes. Each HYPE SLC class represents a unique combination of physical properties such as landuse, soil type, elevation, soil depth, and/or crop/vegetation cover. Routines are included in HYPE for most major land surface and subsurface processes including: snow/ice accumulation and melting, evapotranspiration, surface and macropore flow, soil moisture, discharge generation, groundwater fluctuation, aquifer recharge/discharge, irrigation, abstractions, and routing through rivers, lakes and reservoirs. HYPE's model routines are controlled through general parameters, which provide a single value across the model domain, as well as through parameters that are dependent upon a specific property (typically soil type or landuse), which allow for spatial variability in the routine's behavior. Many water quality constituents can be simulated in HYPE, and these include nitrogen, phosphorus, organic carbon, sediment, silica, general constituent, and water temperature. HYPE simulations are typically performed using a daily time step, but the model has also been applied with 1-hour and 6-hour time steps. Model outputs are typically aggregated across the subbasin area or are generated for the subbasin outlet point with an option to select outputs on the SLC level.

2.2. Model Routines

Complete documentation of all available HYPE's model routines can be found in SMHI (2023). Here we highlight the model routines and options relevant to the setup and calibration of E-HYPE4, specifically E-HYPE version 4.1.3. The version of the HYPE model code used throughout this study was version 5.18.0.

HYPE includes routines to simulate both snow and glacier processes. In E-HYPE4, snowfall and snowmelt processes are simulated using the degree-day method. The fraction of precipitation falling as snow is calculated according to a threshold temperature that can vary with landuse. Mixed rain and snow precipitation occur when the air temperature is within a specified interval of the threshold temperature. Snow melting is proportional to temperature when the temperature is greater than a landuse dependent threshold temperature, and the potential snow that melts can be tuned with an additional landuse dependent parameter. Glacier SLCs consist of two parts: a glacier and a snowfield. Individually, the glacier and snowfield have variable areas, but the total area is always constant. Glacier melting is proportional to temperature.

In HYPE, the local runoff from a subbasin is calculated as the sum of runoff from the land surface (infiltration excess overland flow), runoff from the individual soil layers (interflow), and runoff through subsurface drainage pipes (Fig. 1). Overland flow is computed

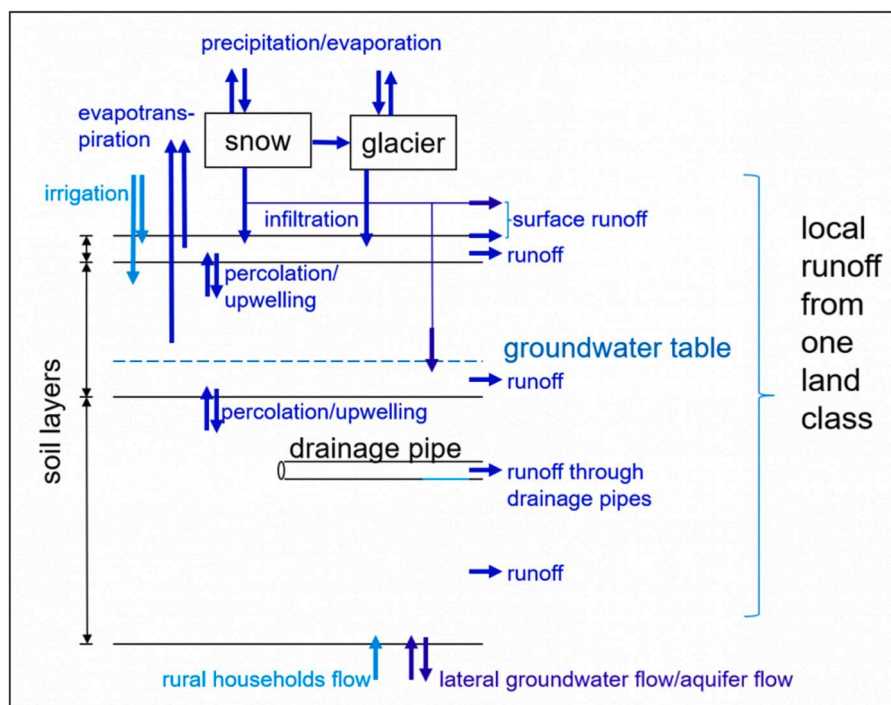


Fig. 1. HYPE model flow paths, from SMHI (2023).

according to runoff coefficients and thresholds for infiltration and soil moisture. In this routine, the total water available for infiltration — the sum of rainfall and snowmelt — is diverted into three flow paths: overland flow to the nearest stream, macropore flow to the soil layer where the groundwater table is currently located, and infiltration to the top soil layer. HYPE models can contain up to three soil layers, but deep aquifers are not simulated in E-HYPE4. Water retention in the soil layers is based on the wilting point, field capacity, and effective porosity. Overland and macropore flows may be generated if the total water available for infiltration exceeds the soil type dependent maximum infiltration rate threshold. In order for macropore flow to be generated, the soil moisture in the top soil layer must also exceed a soil type dependent soil moisture threshold. If these conditions are met, then the overland and macropore flows are calculated as fractions of the total available water exceeding the maximum infiltration rate according to runoff coefficients specified for overland flow and macropore flow, respectively. Total infiltration is then calculated by subtracting the overland and macropore flows from the sum of rainfall and snowmelt. Percolation from an upper to lower soil layer occurs when the soil moisture in the upper layer is greater than the layer's field capacity, and the amount of percolation that occurs is limited by model parameters specifying the maximum percolation rates between the soil layers. The water table is found in the lowest soil layer that is not completely saturated, and the ground water level is estimated according to the level where the soil pores are filled. Runoff from the soil zone occurs when the soil moisture exceeds the field capacity for a soil layer, and runoff through the subsurface drainage pipes occurs when the water table rises above the pipe's depth. If the water table exceeds the surface of the uppermost soil layer, saturation overland flow occurs, and runoff is removed from the top soil layer. In E-HYPE4, runoff and evapotranspiration are calculated before the infiltration and percolation processes in order to slow the response of runoff from the individual soil layers.

Potential evapotranspiration (PET) is simulated in E-HYPE4 using a Modified Jensen-Haise/McGuinness model. This model provides a balance between physical basis and input data requirement, and it was found to provide results that were as efficient as more complex models such as the Penman model and its variants (Oudin et al., 2005). PET in HYPE's Modified Jensen-Haise/McGuinness model is calculated as a function of temperature, crop cover, extraterrestrial radiation, and latent heat of vaporization. Crop cover is tied to the specified landuse while extraterrestrial radiation and latent heat of vaporization are estimated according to latitude and temperature, respectively. Actual evapotranspiration is equal to PET if the soil moisture exceeds a large proportion (HYPE's lp parameter) of the soil field capacity but decreases linearly to zero when the soil moisture is below the wilting point.

HYPE also includes a routine to simulate irrigation. In this routine, the irrigation water demand is first calculated for each irrigated landuse class for each day of the simulation. For submerged crops (e.g. paddy rice), this water demand is calculated to satisfy a target water depth above the soil surface (Wisser et al., 2008). In contrast, for non-submerged crops, the water demand is calculated based on the FAO-56 crop coefficient methods (Allen et al., 1998). Next, HYPE withdraws the demanded water from defined irrigation sources (e.g. lakes, rivers, and/or deep aquifers in the model domain). Finally, the demanded water is applied to the irrigated landuse classes while accounting for water losses between demand, withdrawal, and application.

HYPE uses a conceptual approach to describe the routing of flow between rivers and lakes. HYPE includes two types of rivers: local rivers and main rivers. Local rivers represent the ditches, streams, and rivers within a subbasin that receive only runoff from the subbasin. In contrast, main rivers represent the rivers that receive water from upstream subbasins. The overland flow and runoff from the soil zone for a subbasin are first routed to the local river according to recession parameters depending on soil type and slope. Then, a portion of the flow in the local river can be routed through a local lake representing all of the lakes within a subbasin that receive only runoff from the subbasin (i.e. no contributions from upstream subbasins). The flow exiting the local river and the outflow from the local lake are routed to the main river where it is added to the flow from any upstream subbasins. Translation and dispersion of the main river flow is simulated using a linear channel in series with a linear reservoir (Elenius and Lindström, 2022). Before being passed to downstream subbasins, the main river flow can be routed through an outlet lake/reservoir for which rating curves and regulations can be specified.

Two alternative methods for simulating soil erosion were tested during the E-HYPE4 calibration: the process-based Morgan-Morgan-Finney (MMF) model and the simpler, index-based HBV model. The MMF model calculates soil erosion as the sum of separate rainfall driven and surface runoff driven components and is intended for areas with detailed information on soils, crops, and other agricultural practices (Morgan et al., 1984). In this model, erosion is dependent on seasonally varying crop/ground cover, where the start of the crop growing season is calculated based on degree days. For permanent vegetation classes (e.g. forests), however, ground and crop cover are constant. The second method for simulating erosion, the HBV model, is best suited for large-scale modeling and calculates soil erosion as a function of rainfall, slope, soil type, and landuse (Lidén, 1999; Lidén et al., 2001).

During the E-HYPE4 calibration, two alternative methods were also tested for simulating the in-stream sedimentation and resuspension of sediments: The flow-dependent Modified Bankfull Flow method and the concentration-dependent Simplified Bagnold Equation. In the Modified Bankfull Flow method (SMHI, 2023), the rate of sedimentation/resuspension is based on an estimated bankfull flow, and sedimentation occurs when flows are less than the bankfull flow while resuspension occurs when flows exceed the bankfull flow. The bankfull flow used in the Modified Bankfull Flow method is estimated as the second highest simulated flow during the past year of the simulation multiplied by a calibration factor, and the amount of sediment that is available for resuspension in this method is limited to the amount of sediment which has been previously deposited during the simulation. In contrast, for the Simplified Bagnold Equation method (Neitsch et al., 2011), the rate of sedimentation/resuspension is based on the maximum sediment transport capacity, and sedimentation in this method occurs when suspended sediment concentrations exceed the maximum transport capacity while resuspension occurs when concentrations are lower than the maximum transport capacity (SMHI, 2023). In the Bagnold Equation method, a parameter is included to allow the resuspension limit to be increased up to the maximum transport capacity to represent a potential bank or channel erosion.

A new model routine was implemented in E-HYPE4 to simulate the reduction in lake/reservoir storage volumes due to the accumulation of sediments over time. In this routine, the mass of deposited sediments is calculated for each time step according to the

inflow sediment concentration, the lake/reservoir area, and the sediment settling velocity (SMHI, 2023). Next, the mass and density of the settled sediments are used to compute the change in storage capacity, which then affects both the discharge and the sediment concentration of the lake/reservoir outflows. For E-HYPE4, a general sediment density was specified for the whole model domain, and all sediments deposited within the reservoirs were trapped in the reservoirs for the duration of the simulation (i.e. no flushing/-dredging of reservoir sediments was simulated).

2.3. Data Sources & Model Setup

A variety of open source datasets were utilized in the development of the routing and subbasin characterization for the E-HYPE4 model setup (Table 1). First, MERIT Hydro (Yamazaki et al., 2019) DEM and flow direction/accumulation grids were used to define the model domain and subbasin delineations. Then, the EU-Hydro dataset (European Union, Copernicus Land Monitoring Service, 2019) was used to define the river network, lake polygons, and outlet points. In total, E-HYPE4 is comprised of 42,629 subbasins with a median subbasin area of 170 km². Subbasin areas were partitioned into 97 SLC classes derived from 21 landuse categories, 8 soil types,

Table 1
Datasets used in the development of E-HYPE4.

Category	Data	Source	Link	Reference
Model Domain & Routing	Model Domain & Subbasin Delineation River Network, Lake Polygons, & Outlet Points	MERIT Hydro	http://hydro.iis.u-tokyo.ac.jp/~yamada/MERIT_Hydro/index.html	(Yamazaki et al., 2019)
		EU-Hydro	https://land.copernicus.eu/imagery-in-situ/eu-hydro/eu-hydro-river-network-database	(European Union, Copernicus Land Monitoring Service, 2019)
Subbasin Description	Landuse	CORINE	https://land.copernicus.eu/pan-european/corine-land-cover/clc-2006	(European Union, Copernicus Land Monitoring Service, 2020)
	Glaciers	ESA CCI RGI	http://maps.elie.ucl.ac.be/CCI/viewer/download.php https://www.glims.org/RGI/	(ESA, 2016) (RGI Consortium, 2017)
	Climate Zones	Köppen-Geiger	http://koeppen-geiger.vu-wien.ac.at/present.htm	(Kottek et al., 2006)
	Soils	ESDB DMSW	https://esdac.jrc.ec.europa.eu/ https://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/faounesco-soil-map-of-the-world/en/ https://ec.europa.eu/eurostat/web/nuts/background	(Panagos, 2006) (FAO, 2003)
	Crops	NUTS Various National Sources		(Eurostat, 2022)
	Water Management	GMIA MIRCA	http://www.fao.org/nr/water/aquastat/irrigationmap/index.stm https://www.uni-frankfurt.de/45218023/MIRCA	(Siebert et al., 2010) (Portmann, 2008; Portmann et al., 2010)
Calibration & Validation Data	Point Sources	EEA E-PRTR	https://www.eea.europa.eu/data-and-maps/data/member-states-reporting-art-7-under-the-european-pollutant-release-and-transfer-register-e-prtr-regulation-23 https://essd.copernicus.org/articles/13/1531/2021/	(EEA, 2012)
	Meteorological Forcing Data	HydroGFD		(Berg et al., 2021)
	Precipitation	MODIS	https://lpdaac.usgs.gov/products/mod16a2v006/	(Mu et al., 2011)
	Potential Evapotranspiration	BHDC EWA	https://www.smhi.se/sgn0102/bhdc/index.htm https://www.bafg.de/GRDC/EN/04_spcldtbss/42_EWA/ewa_node.html	
	Discharge	GRDC IMGW SMHI Various National Sources	https://portal.grdc.bafg.de/applications/public.html?publicuser=PublicUser#dataDownload/Home https://www.imgw.pl/ http://vattenwebb.smhi.se/station/	
	Sediment	GEMStat	https://gemstat.org/data/data-portal/	(United Nations Environment Programme, 2018)
		GLORICH	https://doi.pangaea.de/10.1594/PANGAEA.902360	(Hartmann et al., 2019, 2014)
		Naiades Various National Sources	https://naiades.eaufrance.fr/	

and 13 crop types. Landuse data was primarily obtained from the CORINE CLC2006 v2020_20u1 landuse dataset (European Union, Copernicus Land Monitoring Service, 2020). However, in areas outside the CORINE dataset coverage (e.g. parts of Belarus, Greece, Moldova, Russia, & Ukraine), landuse data was obtained from the lower resolution ESA CCI Land Cover v1.6.1 2010 epoch dataset (ESA, 2016). The 21 E-HYPE4 landuse categories were generated by merging the combined landuse data with RGI 6.0 glacier areas (RGI Consortium, 2017), Köppen-Geiger climate zones (Kottek et al., 2006), and GMIA/MIRCA irrigated areas (Portmann, 2008; Portmann et al., 2010; Siebert et al., 2010). Soils data was obtained from the European Soil Database (ESDB) (Panagos, 2006) and Digital Soil Map of the World (DSMW) datasets (FAO, 2003). The E-HYPE4 crop description was produced by merging crop data from various national sources with Nomenclature of Territorial Units for Statistics (NUTS) regions (Eurostat, 2022). Irrigated areas and data on monthly crop growing areas, crop calendars, and cropping periods were obtained from the GMIA (Siebert et al., 2010) and MIRCA (Portmann, 2008; Portmann et al., 2010) datasets. Data on the locations of European industrial facilities and their respective releases were obtained from the E-PTTR v3.3 dataset (EEA, 2012).

The E-HYPE4 model setup was run using meteorological forcing data from HydroGFD3.2 (Berg et al., 2021). The HydroGFD3.2 data set provides bias-adjusted ERA5 reanalysis precipitation and temperature (minimum, mean, & maximum) data globally at a 0.25° horizontal resolution and daily time step. Currently, HydroGFD3.2 is available for the period of 1979 until the present - 5 days. Data from 1979 to 2020 was used for model calibration and validation. Daily precipitation and temperature from HydroGFD were assigned to each E-HYPE4 subbasin via weighted averaging of the overlapping subbasin/grid areas.

Observed potential evapotranspiration, discharge, and sediment concentration data were obtained from a variety of different sources (Table 1). Long-term average potential evapotranspiration was computed for 2671 subbasins using MODIS data from 2000 to 2012 (Mu et al., 2011). Observed discharge and sediment timeseries were assigned to the E-HYPE4 subbasins according to a process which considered both the spatial proximity of the gauging station to the subbasin outlet and the quality of the data set (e.g. presence of gaps in the record). Overall, daily discharge data were available for 2702 subbasins for 1961–2014. Sediment concentration data were available for 1029 subbasins for 1964–2018 with typical sampling frequencies for the observation stations of once per week or once per month. It is worth noting, however, that discharge and sediment concentration timeseries were not available for the complete periods for all stations. Model calibration and validation was performed using a subset of the available observations (See Section 2.4.3).

2.4. Model calibration

The E-HYPE4 calibration was performed by running the model using ensembles of pre-generated parameter sets (Section 2.4.2) and selecting the parameter sets and model routines which optimized the model performance for discharge, long-term average evapotranspiration, and sediment concentrations. Ideally, all of the model parameters controlling discharge and sediment simulations would have been calibrated simultaneously. However, the number of candidate parameter sets required to ensure full coverage of the parameter space for so many individual parameters and model routine options would have been too computationally expensive. Thus,

Table 2
Discharge Calibration Parameters.

Process	Parameter	Type	Units	Min	Max
Precipitation	Correction parameter for precipitation (pcelevadd)	General	(100 m) ⁻¹	0	5
Temperature	Parameter for temperature elevation dependence (tcelevadd)	General	°C (100 m)–1	0	0.7
Evapotranspiration	Correction parameter for temperature (tempcorr)	General	°C	-0.2	0.2
	Factor for calculating soil water limit for potential evapotranspiration (lp)	General	—	0.8	1
Snow	Crop coefficient for potential evapotranspiration models (kc)	Landuse	—	0.6	1.6
	Threshold temperature for snow melt, snow density, and evapotranspiration (ttmp)	Landuse	°C	-5	5
Water holding	Melting parameter for snow (cmelt)	Landuse	mm °C ⁻¹ d ⁻¹	-4	4
	Effective porosity fraction (wcep)	Soil	—	0.01	0.8
	Fraction of soil available for evapotranspiration but not for runoff (wcfc)	Soil	—	0.01	0.6
Percolation	Wilting point fraction (wcwp)	Soil	—	0.01	0.5
	Maximum percolation capacity between soil layers 1 and 2 (mperc1) and layers 2 and 3 (mperc2)	Soil	mm d ⁻¹	5	120
Recession	Recession coefficient for top soil layer (rrcs1)	Soil	d ⁻¹	0.001	0.6
	Recession coefficient for bottom soil layer (rrcs2)	Soil	d ⁻¹	0.001	0.6
	Recession coefficient for slope dependence (rrcs3)	General	d ⁻¹ % ⁻¹	0	0.1
Surface runoff	Recession coefficient for surface runoff (srrcs)	Landuse	d ⁻¹	0.01	0.7
	Fraction for surface runoff (srrate)	Soil	—	0.05	0.5
	Fraction for macropore flow (macrate)	Soil	—	0.05	0.5
	Threshold for macropore flow (mactrinf)	Soil	mm d ⁻¹	0.0007	100
Flood speed	Threshold soil water for macropore flow and surface runoff (mactrsm)	Soil	—	0	1
	Celerity of flood in watercourse (rivvel)	General	m s ⁻¹	0.5	1.50
Lake Rating Curves	Fraction of delay in the watercourse which also causes damping (damp)	General	—	0.4	0.9
	Parameter of rating curve for lake outflow (gratk)	General	—	1	2
Sediments	Parameter of rating curve for lake outflow (gratp)	General	—	1	100
	Exponent for sedimentation/resuspension in watercourses (sedexp)	General	—	0.05	5
	Sedimentation velocity of suspended sediments in lakes/reservoirs (sedss)	General	m d ⁻¹	0.01	0.1
	Soil cohesion (soilcoh)	Soil	kPa	0.01	20
	Soil erodibility (soilerod)	Soil	g J ⁻¹	0.01	2

calibration was conducted using a two-stage approach.

In the first calibration stage, hereinafter referred to as the “discharge calibration”, a multi-objective calibration was performed for discharge, evapotranspiration, and sediments using an ensemble of 20,000 candidate parameter sets based on 24 parameters controlling discharge and 4 parameters controlling soil erosion and lake/reservoir siltation (Table 2). The goal of this stage of the calibration was to calibrate discharge while minimizing the biases in sediment concentrations. Initial values for the uncalibrated sediment routine parameters were obtained from other HYPE model setups which were calibrated for sediments. Parameterization for regulations, dams/reservoirs, and lakes were taken from the previous E-HYPE3 model (Hundecha et al., 2016), for which extensive calibration of the parameters controlling the lake outflows and dam regulations was performed; in E-HYPE3, general parameters were calibrated for a set of subbasins with significant proportions of lakes/regulation, and additional parameters were individually tuned using data from the nearest downstream flow gauge. The discharge calibration was performed for the period of 1991–2010 using 193 stations for calibration and 1465 independent stations for validation. Section 2.4.3 describes the selection procedure for the calibration stations, and a map of all calibration and validation stations is provided in Fig. 1. A spin-up period of 1981–1991 was included during the simulations to allow the model states to stabilize. The individual steps of the discharge calibration procedure are described in Section 2.4.4.

In the second calibration stage, hereinafter referred to as the “sediment calibration”, four combinations of soil erosion and in-stream sedimentation/resuspension routines were run using an ensemble of 5000 candidate parameter sets, resulting in a total of 20,000 unique model setups (4 routine combinations x 5000 parameter sets). The candidate parameter sets were based on 23 parameters controlling soil erosion, in-stream sedimentation/resuspension, and lake/reservoir siltation (Table 3). The goal of this stage of the calibration was to evaluate the different sediment routines and calibrate the sediment fluxes, and a multi-objective calibration was performed for discharge, evapotranspiration, and sediments. The simulation period for the sediment calibration was 2001–2010 with a spin-up period of 1981–2001 to allow the model states to stabilize. The sediment calibration was performed using 131 stations for calibration and 802 independent stations for validation. The selection procedure for the calibration stations is described in Section 2.4.3, and a map of the calibration and validation station locations is provided in Fig. 1. The individual steps of the sediment calibration process are detailed in Section 2.4.5.

2.4.1. Performance criteria

Model performance was assessed during calibration using the Kling-Gupta Efficiency (KGE) model fit statistic. The KGE statistic was

Table 3
Sediment Calibration Parameters.

Process	Parameter	Type	Units	Min	Max
General Soil Erosion	Exponent for delay of suspended sediment from surface runoff and tile drains (pprelexp)	General	—	0	10
	Parameter for delay of suspended sediment from surface runoff and tile drains (pprelmax)	General	mm d ⁻¹	0.001	50
MMF Erosion Model	Time constant of removal of eroded sediment (eroddecay)	General	d ⁻¹	0.01	5
	Exponent for calculation of soil erosion caused by surface runoff (sroexp)	General	—	0	2
	Soil cohesion (soilcoh)	Soil	kPa	0	10
	Soil erodibility (soilerod)	Soil	g J ⁻¹	0	5
HBV Erosion Model	Scaling of subbasin erosion index (erodindex)	General	—	0.01	10
	Erosion precipitation dependent factor (erodexp)	General	—	0	5
	Erosion slope dependent factor (erodslope)	General	—	0	5
	Erosion soil type dependent factor (erodsoil)	Soil	—	0.1	3
	Erosion landuse erosion factor (erodluse)	Landuse	—	0.1	3
Modified Bankfull Flow Sedimentation/Resuspension	Exponent for sedimentation/resuspension in watercourses (sedexp)	General	—	0	2
	Bankfull flow adjustment factor (qbank)	General	—	0	1
Simplified Bagnold Equation Sedimentation/Resuspension	Exponent to calculate river velocity with hydraulic geometry (hygeomk)	General	—	0.001	1
	Rate to calculate river velocity with hydraulic geometry (hygeommm)	General	—	0.001	1
	Peak channel velocity adjustment factor (vpeak)	General	—	0.01	10
	Coefficient for maximum suspended sediment concentration (suspcnss)	General	kg L ⁻¹ (m/s) ⁻¹	0.0001	0.01
	Exponent for maximum suspended sediment concentration (suspcpss)	General	—	1	2
	River channel erodibility/vegetation cover modification factor (suspch)	General	—	0.01	1
	Fraction of sediment in excess of temporary pool that can be resuspended (addsusp)	General	—	0	1
	Lake/reservoir sediment density lseddens	General	kg m ⁻³	400	1600
	Sedimentation velocity of suspended sediments in lakes/reservoirs (sedss)	General	m d ⁻¹	0	100
Lake/Reservoir Siltation	Reduction in suspended sediment concentration for calculation of lake/reservoir sedimentation (limsedss)	General	mg L ⁻¹	0	10

developed by (Gupta et al., 2009) to facilitate the analysis of the relative importance of the correlation, bias, and variability ratio components of the popular Nash-Sutcliffe Efficiency (NSE) performance criterion. Analysis of the individual KGE components can provide a more complete understanding of the model performance, and weighting of the components can allow modelers to fine tune the performance according to the model's desired purpose. During calibration, the KGE bias component was calculated as a relative error in order to assess if the model was over or underestimating the observed values. KGE ranges from negative infinity to one with $KGE = 1$ indicating a perfect agreement between observed and simulated values. The optimal values for the correlation, relative error, and variability ratio KGE components are 1, 0, and 1, respectively. Values of KGE in the range $-0.41 < KGE \leq 1$ indicate that the model improves upon using the mean observed value as a benchmark predictor (Knoben et al., 2019). Pearson correlation coefficients > 0.7 were considered to demonstrate high positive correlations, and relative errors within $\pm 25\%$ were considered satisfactory (Moriasi et al., 2007).

2.4.2. Parameter sampling

The Latin Hypercube sampling method within the SAFE toolbox (Pianosi et al., 2015) was used to generate the ensembles of candidate parameter sets used for the model calibration. This method ensures a very efficient coverage of the parameter space and can be used to specify parameter distributions and relationships between parameters (i.e. value for one class must be greater than value for another class). In addition, unlike HYPE's in-built Monte-Carlo Differential Evolution Markov Chain autocalibration methods (SMHI, 2023) which start from a random seed and may converge on slightly different optimal parameter values when run multiple times, using pre-generated parameter ensembles ensured that the results would be identical if the calibration process were repeated. Furthermore, running the parameter ensemble and saving the results from each model run meant that alternative objective functions or calibration/validation stations could be used without re-running the model.

For the discharge calibration, the ensemble of 20,000 candidate parameter sets was generated from 28 calibration parameters (Table 2) controlling discharge, soil erosion, and lake/reservoir siltation. Of the 28 parameters, 11 were general parameters, 13 were tied to soil type, and 4 were tied to landuse class. Values were generated for each of the 8 soil types and 21 landuse categories; thus, the total number of optimized parameters was 199 (11 general parameters + 13 soil parameters $\times$ 8 soil types + 4 landuse parameters $\times$ 21 landuse classes). The search range for each parameter was pre-defined subjectively based on prior experience, and relationships were

Table 4

Representative Gauged Basin (RGB) selection for the discharge calibration. A total of 226 unique RGBs were selected: 97 subbasins for general model parameters, 159 for soil-specific parameters, and 216 for landuse-specific parameters (Several subbasins were included in multiple RGB categories). Dashes in the Count column indicate the soils, landuses, and crops for which there were no representative subbasins.

General	Percent of Model Domain	Maximum Upstream Area (km ²)	Upstream Regulation Allowed	Count
Unregulated Subbasins	—	1000	False	97
Soil Type	Percent of Model Domain	Minimum Upstream Area Fraction	Upstream Regulation Allowed	Count
Coarse	10.37	0.4	False	19
Medium	30.22	0.4	False	53
Fine	18.10	0.4	False	11
Organic	6.72	0.4	False	5
No Texture (i.e. Rock Lakes)	6.17	0.2	False	—
Shallow	17.18	0.4	False	27
Moraine	11.15	0.4	False	55
Fine for Rice Crops	0.09	0.2	False	—
Landuse Class	Percent of Model Domain	Minimum Upstream Area Fraction	Upstream Regulation Allowed	Count
Lake	4.48	0.2	False	—
Glacier	0.06	0.2	False	—
Urban	1.34	0.2	False	—
Semi-Urban	2.67	0.2	True	12
Deciduous Forest	6.37	0.2	False	27
Coniferous Forest	15.81	0.4	False	54
Mixed Forest	6.38	0.2	False	11
Agriculture	28.45	0.4	False	28
Irrigated Agriculture	2.83	0.2	False	—
Perennial Agriculture	1.46	0.2	True	14
Pasture	5.59	0.4	False	5
Open with Vegetation	6.50	0.2	False	65
Open without Vegetation	6.16	0.4	False	3
Wetland	3.74	0.6	False	—
River	0.29	0.2	True	—
Mediterranean Open	3.29	0.4	True	26
Mediterranean Deciduous Forest	1.46	0.2	False	2
Mediterranean Coniferous Forest	0.89	0.2	False	—
Mediterranean Mixed Forest	0.42	0.4	True	2
Northern Deciduous Forest	1.80	0.4	True	—
Irrigated Pasture	0.01	0.2	True	—

specified between the coarse, medium, and fine texture soil types for various soil parameters to ensure that, for example, coarser texture soils had higher maximum percolation capacities than finer texture soils.

In the sediment calibration, the ensemble of 5000 candidate parameter sets was generated from 23 parameters (Table 3) used in the general soil erosion and lake/reservoir siltation processes, the two soil erosion models, and the two-stream sedimentation/resuspension methods (Section 2.2). Of the 23 parameters, 19 were general parameters, 3 were tied to soil type, and 1 was tied to landuse class. Because some of the soil and landuse classes were not overly prevalent within the model domain, parameter values were only

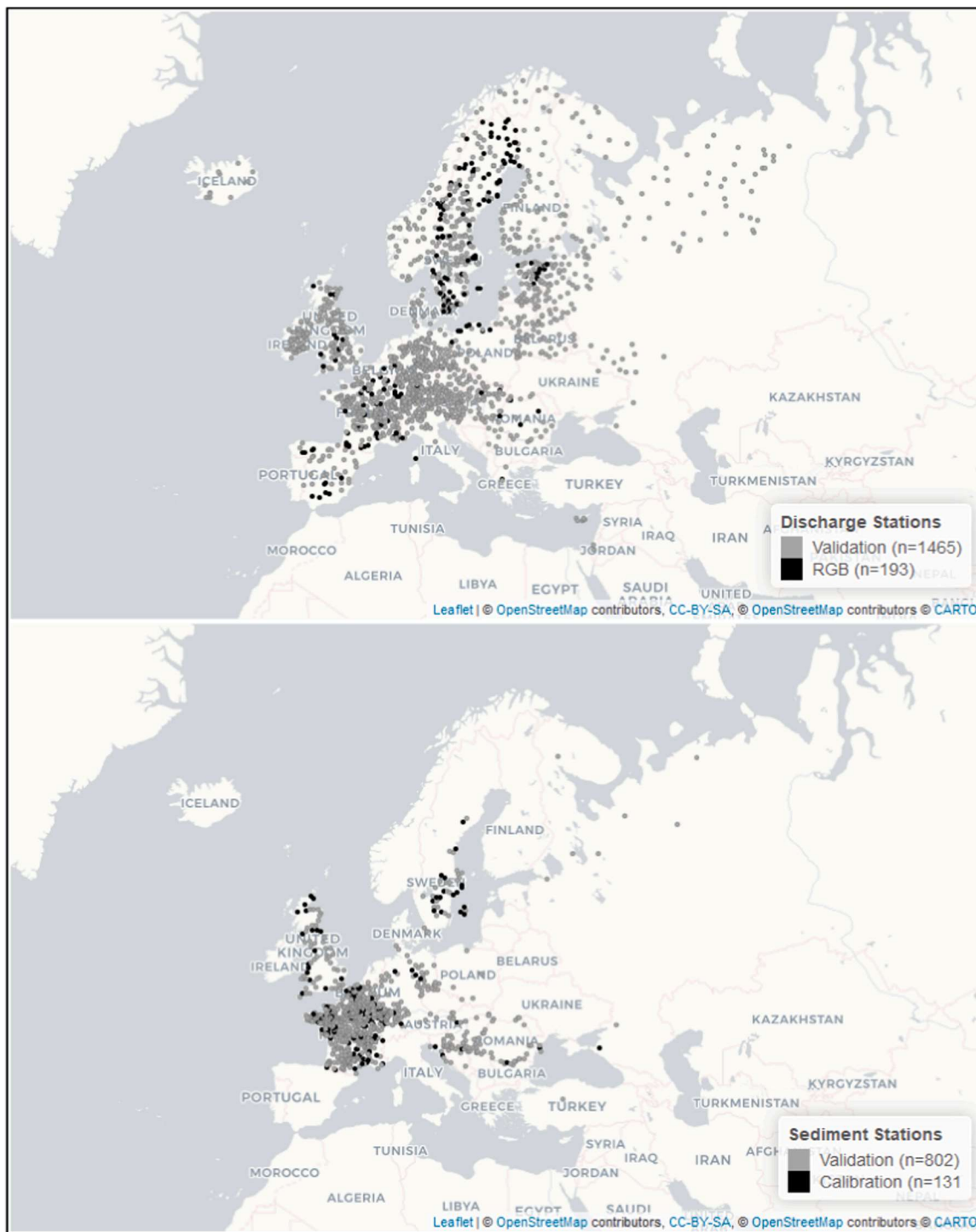


Fig. 2. Locations of calibration vs. validation stations for discharge and sediments.

generated for the nine most dominant landuse classes and the five most dominant soil types according to subbasin area fraction; thus, the total number of optimized parameters was 43 (19 general parameters + 3 soil parameters x 5 soil types + 1 landuse parameter x 9 landuse classes). As in the discharge calibration, the search range for each sediment parameter was pre-defined subjectively based on prior experience. During sampling, relationships were specified such that soil erodibility was greater for coarser texture soil types than for finer texture soils. Relationships were also specified to ensure that soil cohesion was greater for finer texture soils than for coarser texture soils. In addition, the parameter ranges for the four parameters that were included in the discharge calibration were updated according to the knowledge gained from the discharge calibration results.

The selection of the calibration ensemble sizes was based on several considerations. First, the ensemble sizes were pragmatic as the model simulations for the full ensembles could be completed with the available computing resources within the target of approximately 24 h. An additional consideration for selecting an ensemble of 5000 parameter sets for the sediment calibration was that, because four combinations of model routines were tested for each parameter set (Section 2.4.5), the total number of model setups evaluated during the sediment calibration ($n = 20,000$) was equal to that of the discharge calibration. The ensemble sizes were also selected to provide ratios between the number of candidate parameter sets to the number of optimized parameters of $\geq 100:1$. The informal rule regarding the number of parameter sets to generate for an effective experiment is that the number of parameter sets should be approximately 10 times the number of optimized parameters (Chapman et al., 1994; Jones et al., 1998; Loeppky et al., 2009). Although using smaller ensembles may have been sufficient, the larger selected ensemble sizes provided additional confidence that the ensembles were large enough to achieve adequate coverage of the sampled space and resolve all possible local minima points.

2.4.3. Catchment classification

The stations used for the discharge calibration were selected through a representative gauged basin analysis in which groups of gauged subbasins with upstream areas dominated by a certain soil type or landuse class were identified (Donnelly et al., 2016; Strömqvist et al., 2012). Calibrating the model to the observations for these representative gauged basins (RGBs) allows for (1) the isolation and tuning of the key processes and characteristics for specific soil types and landuse classes and (2) the evaluation of the model performance separately for individual soil types and landuse classes. In total, there were 1658 candidate subbasins with discharge observations. For each of these subbasins, a check for the presence of upstream regulations was performed, and the upstream area fractions for the 8 soil types and 21 landuse classes were calculated. When establishing the selection criteria for the RGBs, there was a tradeoff between selecting candidates with no upstream regulations with high upstream area fractions (i.e. dominated by a soil type or landuse) and selecting enough sites to provide coverage across the model domain. Ultimately, the criteria used for the RGB selection (Table 4) attempted to provide a balance between these two conditions. All selected RGBs included a minimum of three years of observations. Overall, 226 unique RGBs were selected, with 97 pristine (non-regulated) subbasins selected to represent general model parameters, 159 selected to represent soil-specific parameters, and 216 selected to represent landuse-specific parameters (Several subbasins were included in multiple RGB categories). The 193 RGBs with discharge observations for the 1991–2010 discharge calibration period were used for model calibration, representing 11% of the model domain area, while the remaining 1465 independent observation stations were used for validation.

Because the available sediment observation stations were not well distributed spatially across the model domain, selection of the sediment calibration stations required a different process from that which was used to select the discharge calibration stations. In total, there were 933 candidate subbasins with sediment concentration data available for the calibration period of 2001–2010. To isolate and tune the sediment processes, sites without upstream lakes/reservoirs were excluded if they had upstream areas less than 100 km² or greater than 500 km², and sites with upstream lakes/reservoirs were excluded if the lake/reservoir was not within 100 km of the subbasin outlet or if less than 80% of the upstream area drained through the lake/reservoir. To avoid biasing the model performance towards regions containing more observation stations, a subset of the remaining candidate sites was then selected to provide a set of subbasins with similar distributions of selected catchment characteristics to those of the full model domain. Relative frequency histograms for the full model domain were generated for 17 catchment characteristics: upstream area fractions for the five calibration soil types and nine calibration landuse classes, mean subbasin slope, total subbasin precipitation, and average annual median sediment concentration. For each characteristic, a random sample of the candidate subbasins was taken. Then, subbasins were selectively removed from the sample in order to adjust the relative frequencies of the histogram bins to match those from the full model domain distribution as closely as possible. The candidates selected as sediment calibration subbasins were those which appeared in the adjusted samples for at least 7 of the 17 catchment characteristics. Overall, 131 subbasins were selected for the sediment calibration, and the remaining 802 independent subbasins were used as validation stations.

2.4.4. Discharge calibration

The E-HYPE4 discharge calibration followed an iterative process in which regional biases in the water balance were corrected while simultaneously optimizing model performance for discharge, evapotranspiration, and sediments. The discharge calibration was divided into three phases: Water Balance Check, Discharge Calibration, and Custom Tuning (Fig. 2).

The Water Balance Check phase (Fig. 2, Steps 1 & 2) was performed to identify and correct for systematic regional volume biases in the water balance. It is well known that precipitation gauges tend to e.g. underestimate the amount of rainfall in mountainous areas (Gentilucci et al., 2021). To correct for these types of errors, the full 20,000 discharge calibration parameter set ensemble (Section 2.4.2) was first run on a submodel consisting of the RGB calibration basins and their upstream drainage areas (Fig. 2, Step 1). Plots were generated to compare the relative errors in discharge with the calibration parameter values to verify that the search range for each parameter was adequate. Then, the results were aggregated, and the individual parameter sets which minimized the relative error between the simulated and observed discharge for each of the 193 calibration stations were identified. This analysis resulted in a

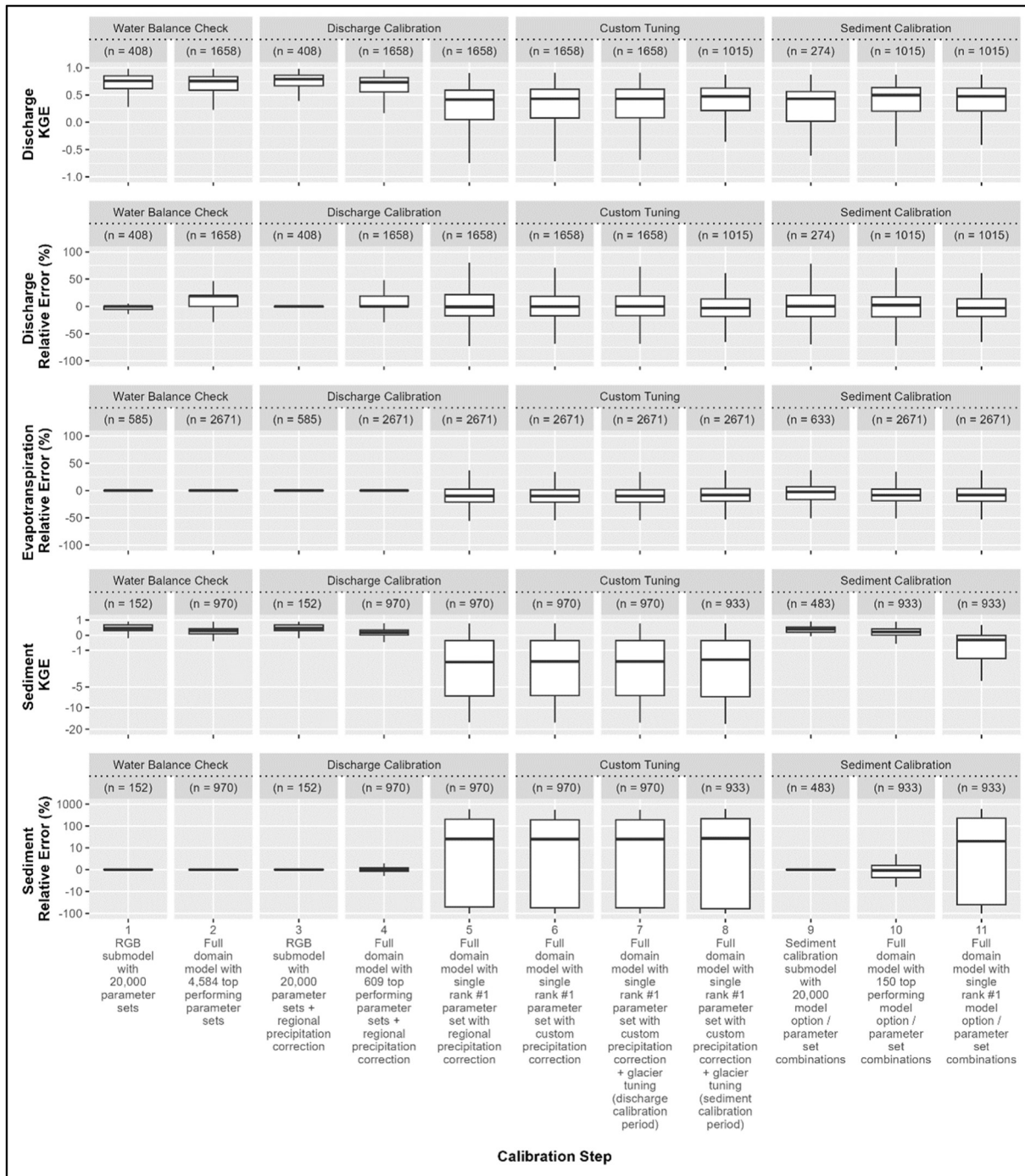


Fig. 3. Maximum achievable model performance during each step of the calibration process for every calibration and validation station included within the model domain for the respective calibration step. Each data point in the boxplots represents a separate station, and the maximum achievable performance describes the highest KGE and the lowest absolute magnitude relative that could be obtained at the station using any single parameter set from the ensemble of candidate parameter sets for the given calibration step. The number of candidate parameter sets for each calibration step is indicated in the x-axis labels. The labels above the boxplots indicates the number of observation stations represented in the boxplots, which varies because the calibration was performed on both the full model domain as well as on submodels containing only the calibration stations and their upstream areas. The simulation period for Steps 1–7 was the discharge calibration period of 1991–2010, whereas the simulation period for Steps 8–11 was the sediment calibration period of 2001–2010. Outliers are hidden from the boxplots.

subgroup of 4584 top-performing parameter sets, and the full domain model was run for each of these parameter sets (Fig. 2, Step 2). Next, the model was assessed for systematic regional volume biases in the water balance by using the SKATER algorithm (Assunção et al., 2006) to cluster the subbasins with discharge observations according to their best achievable relative errors in discharge (i.e. relative error with lowest magnitude) from the 4584 model runs. Finally, preliminary regional precipitation correction factors were applied to each cluster to correct the volume biases. These regional correction factors were selected according to the median relative error in discharge for the individual clusters.

During the Discharge Calibration phase (Fig. 2, Steps 3–5), a single optimal parameter set for the whole model domain was identified from the full 20,000 discharge calibration parameter set ensemble. First, the full 20,000-member ensemble was rerun using the RGB submodel with the regional precipitation correction factors applied (Fig. 2, Step 3). Then, a pareto front analysis was conducted on the performance for the calibration stations with criteria to maximize the median KGE for discharge while minimizing the magnitude of the median relative error for both discharge and evapotranspiration. In total, the pareto front analysis identified a set of 609 top performing parameter sets. Each of these 609 parameter sets was then run for the full domain model with the regional precipitation correction factors applied (Fig. 2, Step 4). Next, the model performance for the validation stations was aggregated, and a pareto front analysis of the validation results was performed to exclude the inferior parameter sets. Criteria for this pareto front analysis were to maximize median KGE for discharge while minimizing the magnitude of the median relative error for discharge, evapotranspiration, and sediments. Identification of the single parameter set providing the optimal performance across the full model domain was performed via a rescaled ranking of the remaining parameter sets according to both calibration and validation performance (KGE & relative error). In this ranking process, the individual model fit statistics were each rescaled to the range 0–1, and then the parameter set with the largest sum of the individual rescaled statistics was considered the optimal parameter set (Fig. 2, Step 5).

The objectives of the Custom Tuning phase (Fig. 2, Steps 6–8) were to: (1) reassess the regional volume biases for the model simulation using the optimal discharge parameter, (2) apply new precipitation correction factors customized for the optimal parameter set, and (3) perform a manual tuning of the glacier parameters. To assess the regional biases in the water balance for the simulation using the optimal parameter set, the preliminary regional precipitation correction factors used during the previous calibration steps were removed from the model, and the model was rerun without any precipitation corrections. Then, the SKATER algorithm (Assunção et al., 2006) was used to cluster the discharge observation stations by their relative error in discharge (Fig. 3), and each of the eight resulting clusters was assigned a name according to the geographical location with the most stations in the cluster. Stations with relative errors greater than 100% when the model was run without the precipitation corrections ($n = 104$) were excluded from the clustering analysis because retaining these stations in the analysis would have resulted in very small clusters (i.e. only one or two stations) and these atypical biases may have indicated, for example, regulations or other anthropogenic factors that were not captured by the model. Precipitation correction factors were selected subjectively for these new clusters according to the median relative error for each cluster; A single, cluster-specific precipitation correction factor (limited to $\pm 20\%$) was applied to all of the subbasins within each respective cluster except for in the Central Europe cluster where no correction was applied. The model was then run for the optimal discharge parameter set with the new precipitation correction factors applied (Fig. 2, Step 6). Finally, after tuning of the glacier parameters, the model was run for both the discharge and sediment calibration periods (Fig. 2, Steps 7 & 8).

2.4.5. Sediment Calibration

The E-HYPE4 sediment calibration (Fig. 2, Steps 9–11) assessed the performance of different combinations of sediment model routines while simultaneously optimizing model performance for discharge, evapotranspiration, and sediments. The hydrological parameters selected during the discharge calibration were not recalibrated during the sediment calibration. Instead, the sediment calibration focused on the model routines and parameters controlling soil erosion, in-stream sedimentation/resuspension, and lake/reservoir siltation. In total, four sets of sediment routines were evaluated, the unique combinations of the two soil erosion models (MMF & HBV) with the two in-stream sedimentation/resuspension routines (Modified Bankfull Flow & Simplified Bagnold Equation).

The first step in the sediment calibration was to run each of the 20,000 sediment calibration ensemble candidates (4 routine combinations x 5000 parameter sets) for a submodel consisting of the 131 sediment calibration subbasins and their upstream drainage areas (Fig. 2, Step 9). Plots were generated to compare the relative errors in sediments with the calibration parameter values to verify that the search range for each parameter was adequate. Then, the results for the calibration stations were aggregated for each of the sediment routine combinations, and pareto front analyses were performed with criteria to maximize the median KGE for sediments while minimizing the magnitude of the relative error for sediments. Based on a pareto front analysis of the performance for the calibration subbasins for each of the sediment routine combinations, a set of 150 top performing candidates was identified.

The second step in the sediment calibration was to run the full domain model for the 150 top performing candidates from the pareto front analysis (Fig. 2, Step 10). The model performance for the validation stations was then aggregated, and a pareto front analysis of the validation results was performed to exclude the inferior parameter sets. Criteria for this pareto front analysis were to maximize the median KGE for discharge and sediments while minimizing the magnitude of the median relative error for discharge, evapotranspiration, and sediments. The model performance for evapotranspiration was deemed to be equally good for all remaining candidate members. Thus, the single candidate providing the optimal performance across the full model domain was identified by conducting a rescaled ranking of the candidates according to their calibration and validation performance (KGE & relative error) for discharge and sediments. In this ranking process, the individual model fit statistics were each rescaled to the range 0–1, and then the parameter set with the largest sum of the individual rescaled statistics was considered the optimal parameter set (Fig. 2, Step 11).

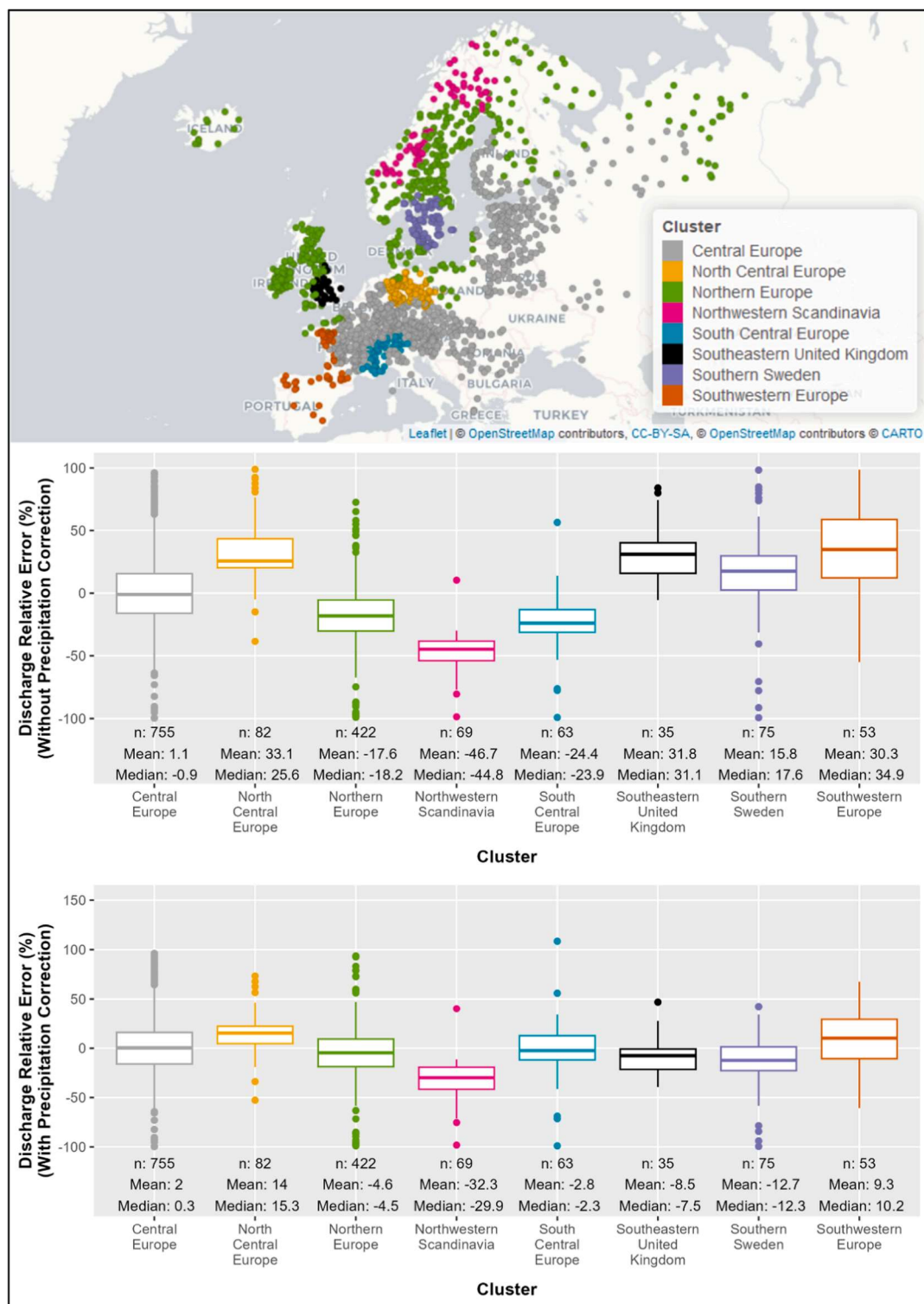


Fig. 4. Clusters of E-HYPE4 discharge observation stations used to select precipitation correction factors to correct regional biases in the water balance for the final selected model. The SKATER algorithm was used to generate clusters according to spatial proximity and relative error in discharge, and each data point represents a separate discharge observation station. Boxplots present the discharge relative errors for the stations within each cluster before and after the precipitation correction factors were applied. Observation stations with discharge relative errors greater than 100% before the precipitation correction factors were applied were excluded from the analysis as these atypical biases may have indicated e.g. regulations or other anthropogenic factors that were not captured by the model.

2.5. Model assessment

2.5.1. Maximum achievable performance

In order to assess the impact of regionalization on the achievable model performance, the maximum achievable model performance was calculated for each calibration and validation station for each step of the calibration process (Fig. 2). The maximum achievable performance for a station was defined as the highest KGE and the lowest magnitude relative error that could be obtained at the station using any single parameter set from the ensemble of candidate parameter sets for a given calibration step. Because the maximum achievable performance for each station could be obtained from any single parameter set, the achievable performance for a station was independent from the achievable performance for any upstream stations. For example, Step 4 of the calibration included 609 candidate parameter sets, so the maximum achievable KGE for a station during this step was the highest KGE value obtained at the station when considering all 609 individual model runs, and each of the stations could obtain their maximum achievable KGE from any of the 609 parameter sets. In contrast, Step 5 of the calibration only included the single optimal discharge parameter set; thus, the maximum achievable KGE for a station during this step was simply the KGE from the single model run, and all of the stations obtained their maximum achievable performance from the same single parameter set.

2.5.2. Overall model performance

The model performance for discharge, evapotranspiration, and sediments using the best global model setup is summarized in Fig. 4, Tables 5 and 6. Throughout the model calibration process (Section 2.4.4 & 2.4.5), the identification of the top performing parameter sets from the calibration ensembles was conducted using median statistics so that the process would not be sensitive to outliers in model performance (i.e. subbasins with extremely good/poor performance). However, mean statistics are also presented in Tables 5 and 6 summarizes the performance of the best global model setup relative to the performance benchmarks described in Section 2.4.1. Model fit statistics for discharge and sediments were calculated for each station from daily data while the model fit statistics for evapotranspiration were based on long-term average values.

2.5.3. Bias by catchment characteristic

In order to identify factors contributing to biases in the discharge and sediment fluxes in the best global model setup, model biases were plotted against catchment characteristics for the observation stations within the seven regions where precipitation correction factors were applied (Fig. 5). Model biases were calculated as a scaled relative error ranging from – 100–100% to facilitate comparison between different catchment characteristics:

$$RE_{scaled} = \frac{Sim - Obs}{Sim + Obs} * 100\%$$

The catchment attributes summarized for each station were the latitude of the centroid of the subbasin, the total upstream catchment area, the mean elevation and slope of the upstream area, and the fraction of the upstream catchment area comprised of various soil types, landuse classes, and crop categories.

3. Results

3.1. Maximum achievable performance

The results presented in Fig. 2 indicate that the degree of regionalization had a significant impact on the achievable model performance. Unsurprisingly, the maximum achievable performance was greatest for the steps which included the most candidate parameter sets and thus provided the greatest potential regionalization (Steps 1, 3, and 9). When individually assessed across the full model domain, each of the 20,000 candidate parameter sets in these calibration steps provided model performance that was $\leq$ that of the single final optimal parameter set identified through the rescaled ranking process (Section 2.4.4 & 2.4.5); however, because each

Table 5

Model performance for the best global model setup. Discharge and evapotranspiration were evaluated for the period 1991–2010, and sediment was evaluated for the period 2001–2010. KGE, correlation, and variability ratio were not calculated for evapotranspiration because the model was evaluated against long term average observed evapotranspiration data.

Objective	Stations	n	KGE		Correlation		Relative Error (%)		Variability Ratio	
			Mean	Median	Mean	Median	Mean	Median	Mean	Median
Discharge	Calibration	193	0.03	0.50	0.69	0.72	14.08	-7.08	1.36	0.95
	Validation	1465	-0.60	0.42	0.64	0.69	55.59	1.06	1.83	1.14
	All	1658	-0.52	0.43	0.65	0.69	50.76	0.01	1.77	1.11
Evapotranspiration	Calibration	222	—	—	—	—	-15.18	-15.99	—	—
	Validation	2449	—	—	—	—	-10.47	-9.52	—	—
	All	2671	—	—	—	—	-10.86	-10.12	—	—
Sediment	Calibration	131	-1.10	-0.30	0.20	0.22	93.30	0.49	0.90	0.42
	Validation	802	-3.18	-0.30	0.31	0.35	300.29	24.29	1.93	0.63
	All	933	-2.89	-0.30	0.30	0.33	271.23	20.11	1.78	0.60

Table 6

Model performance for the best global model setup relative to performance benchmarks. Discharge and evapotranspiration were evaluated for the period 1991–2010, and sediment was evaluated for the period 2001–2010. KGE and correlation were not calculated for evapotranspiration because the model was evaluated against long term average observed evapotranspiration data.

Objective	Stations	n	Percent of Stations With:		
			KGE > -0.41	Correlation > 0.7	Relative Error Within +/- 25%
Discharge	Calibration	193	90.7	53.4	60.6
	Validation	1465	87.1	47.6	63.7
	All	1658	87.5	48.3	63.3
Evapotranspiration	Calibration	222	—	—	70.7
	Validation	2449	—	—	78.5
	All	2671	—	—	77.8
Sediment	Calibration	131	60.3	1.5	17.6
	Validation	802	56.7	2.4	18.1
	All	933	57.2	2.3	18.0

station's maximum achievable performance could be obtained from any of the 20,000 candidate parameter sets, the overall maximum achievable performance for these calibration steps was quite high. Conversely, the greatest decreases in achievable model performance occurred between Steps 4 & 5 and Steps 10 & 11, where the calibration transitioned from using an ensemble of parameter sets to using a single parameter set across the full model domain (i.e. no regionalization). These results indicate that no single parameter set can provide exceptional performance across the full model domain and that there are some processes and/or catchment descriptors which cannot be captured by a single model setup.

Inspection of Fig. 2 reveals several additional key findings regarding the maximum achievable performance throughout the various calibration steps. Introduction of the “regional” precipitation correction factors during the initial stages of the model calibration (See Section 2.4.4) provided substantial improvements in the maximum achievable performance for discharge KGE and relative error (Compare Steps 1 & 3). However, switching from the regional precipitation corrections to the custom parameter set specific precipitation corrections applied to the single best global model setup provided less significant gains in performance (Compare Steps 5 & 6). Slight increases in model performances were also accomplished with the tuning of the glacier parameters, but these gains were less apparent because the proportion of observation sites with upstream glaciers was small in comparison to the total number of observation sites (Compare Steps 6 and 7). Before selection of the optimal discharge parameter set (Step 5), small relative errors (+/- 0.05%) in evapotranspiration were possible for most subbasins; however, the biases in evapotranspiration were not substantially affected further during subsequent calibration (Steps 6–11). Finally, calibration of the model during the Sediment Calibration phase provided notable improvements in the maximum achievable KGE for sediments but less substantial improvements in the achievable relative error for sediments (Compare Steps 8 & 11).

3.2. Regional precipitation corrections

The eight regional precipitation correction factor clusters used in the best global model setup are displayed in Fig. 3. Overall, the number of stations in each cluster ranged from 35 to 755, and the median relative errors for the clusters ranged from -44.8–34.9%. However, the median relative error for the largest cluster (Central Europe) was only -0.9%. The addition of the precipitation corrections improved the regional biases in the water balance across the model domain (Fig. 3). Although no correction factor was applied to the Central European cluster, the mean and median relative error changed slightly because some of the subbasins in the cluster had upstream basins that belonged to a different cluster in which a correction factor was applied.

3.3. Sediment calibration ensemble

Fig. 6 presents the model performance for each of the 20,000 sediment calibration ensemble members that were run for the submodel consisting of the sediment calibration subbasins and their upstream areas. Overall, there was no apparent difference in model performances between the HBV and MMF soil erosion models. However, the ensemble members which used the Modified Bankfull Flow routine for stream sedimentation/resuspension typically provided higher median KGE values than those which used the Simplified Bagnold Equation. Despite this, correlations between the simulated and observed sediment concentrations were higher for the members using the Bagnold equation than for those using the Modified Bankfull Flow routine. Further inspection of the results shows that, for a given KGE value, the members using the Bagnold equation often had greater relative errors and lower variability ratios than those using the Modified Bankfull Flow routine (the Bagnold points appear farther to the right than the Modified Bankfull points in the Relative Error plot and farther to the left than the Modified Bankfull points in the Variability Ratio plot). For each of the four combinations of erosion models and sedimentation/resuspension routines, median KGE values decreased significantly when the median relative error values exceeded 100% or when the variability ratio exceeded 2. In general, the MMF soil erosion model with the Modified Bankfull Flow routine produced the highest KGE values, but the combination of routines significantly underestimated the sediment concentrations for most parameter sets; half of the members with this combination of options had a median relative error between -100% and -94.9%. Ultimately, the combination of model options which provided the highest ranking in the selection of the best global model setup (Section 2.4.5) was the MMF soil erosion model with the Bagnold equation for stream sedimentation/

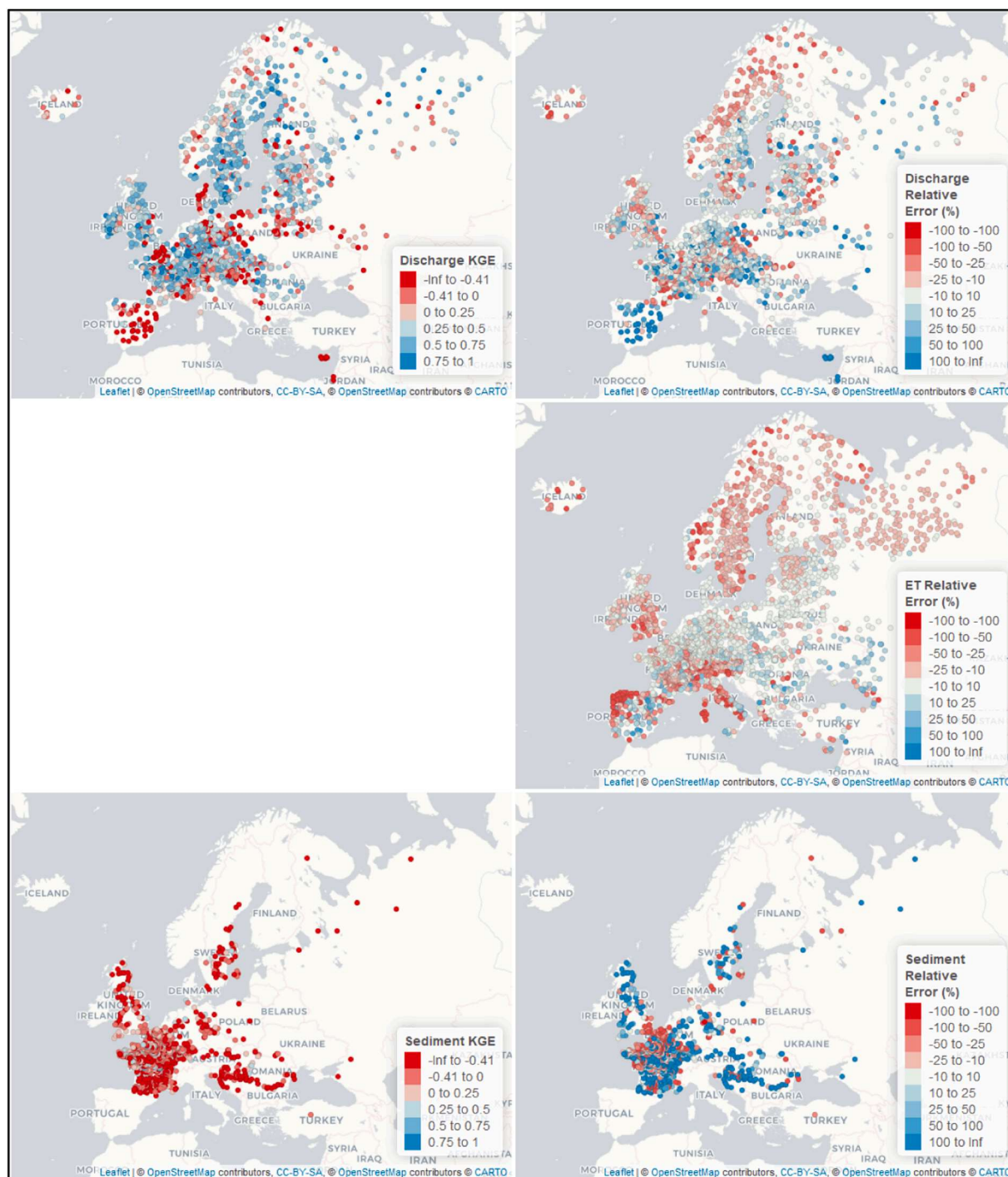


Fig. 5. Model performance for the best global model setup. Discharge and evapotranspiration were evaluated for the period 1991–2010, and sediment was evaluated for the period 2001–2010.

resuspension.

3.4. Overall model performance

In general, the median discharge KGE, correlation, relative error, and variability ratios for the best global model setup were similar between the calibration and validation stations (Table 5). Overall, the median discharge KGE was 0.50 for the calibration stations and

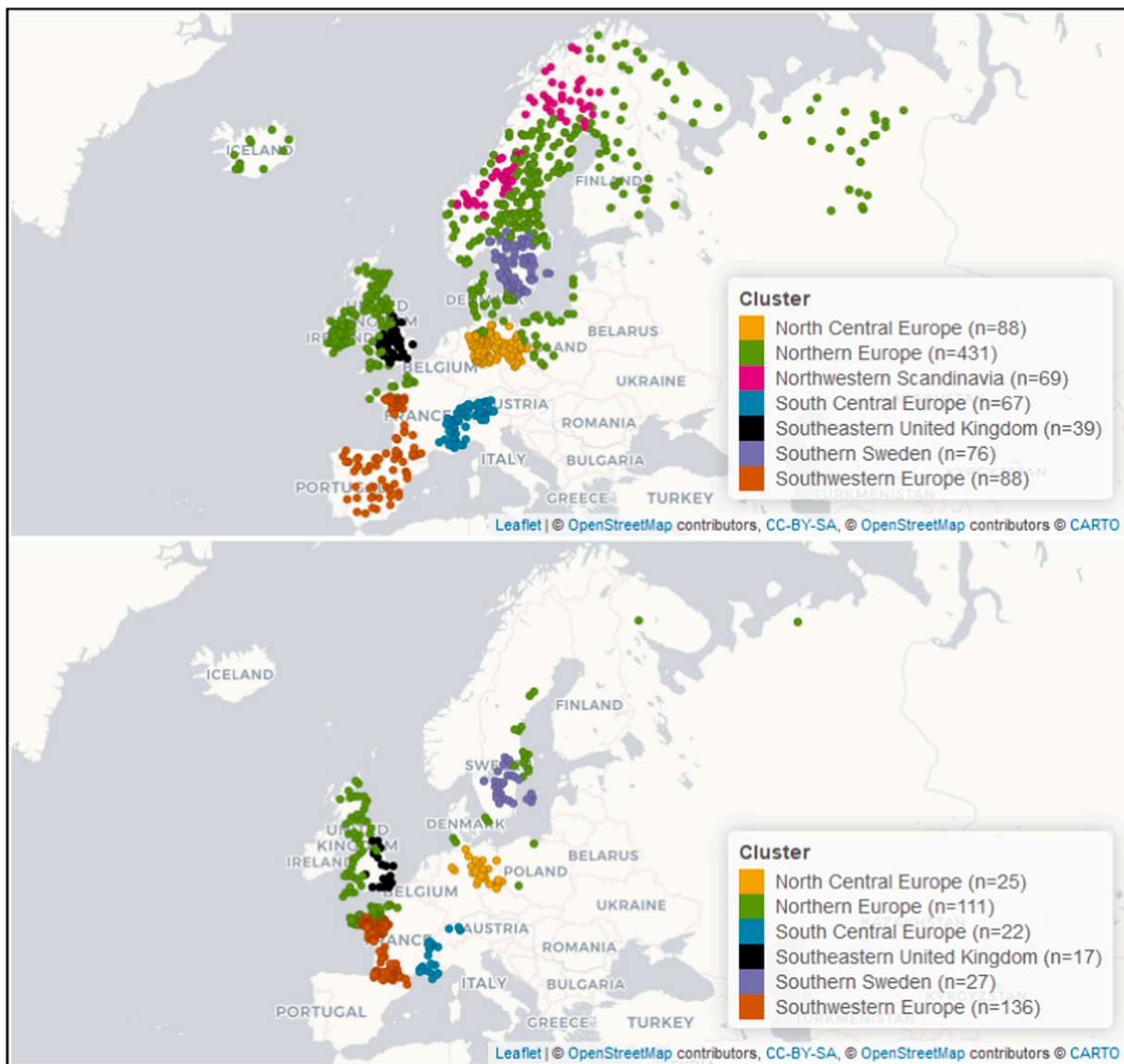


Fig. 6. Locations of discharge observation stations (top) and sediment observation stations (bottom) within precipitation correction clusters. The number of stations within each cluster is indicated in the map legends.

0.42 for the validation stations. Spatially, discharge performance was best for Southern Sweden, the Southeastern United Kingdom, and Central Europe, whereas performance was poorer for Southwestern Europe, where the model significantly overestimated discharges, and Northwestern Scandinavia, where the model significantly underestimated discharges (Fig. 4). The mean discharge biases for the clusters are non-zero because a single cluster-specific precipitation correction factor was applied to all subbasins within the respective cluster (Section 2.4.4) and because the individual observation stations had varying biases in discharge (before the precipitation correction was applied). It is also worth noting that the clustering did not account for the routing between subbasins, and thus, subbasins may have upstream subbasins located in a different cluster in which different precipitation correction factors were applied. The presence of individual subbasins with extreme biases in discharge is evident when examining the mean relative error values. Although the calibration and validation sites had similar median relative errors in discharge (-7.08% and 1.06% , respectively), the mean relative errors were much higher for the validation stations (55.59%) than for the calibration stations (14.08% , Table 5). Despite this, 63.3% of the stations had relative errors within $\pm 25\%$, and 48.3% of the stations had correlation coefficients > 0.7 . Furthermore, 90.7% of the calibration stations and 87.1% of the validation stations had discharge KGE values > -0.41 (Table 6), indicating that the model improved upon using the mean observed value as a benchmark predictor for the majority of the stations.

Biases in evapotranspiration in the best global model setup were also quite similar between the calibration and validation stations. Overall, the median relative errors in evapotranspiration were -16.0% for the calibration stations and -9.52% for the validation

stations (Table 5). In general, the model tended to underestimate evapotranspiration for higher latitudes and overestimate evapotranspiration for lower latitudes (Fig. 4). There were few subbasins with extreme biases in evapotranspiration, and the mean relative errors in evapotranspiration for the calibration and validation stations were -15.18% and -10.47% , respectively. Altogether, 70.7% and 78.5% of the calibration and validation stations, respectively, had relative errors in evapotranspiration within $\pm 25\%$ (Table 6).

Model performance for sediments was limited after the single final parameter set was used across the full model domain (Fig. 2, Step 11), and, in general, the best global model setup overestimated the observed sediment concentrations across much of the model domain. The median relative error for all stations was 20.1%, but there were many instances in which the sediment fluxes were significantly overestimated, and the mean relative error for all stations was 271.23% (Table 5). For perspective, other studies have considered suspended sediment simulations to be good for errors in the mean concentration with magnitudes in the range of 20–30% and very good for magnitudes $\leq 20\%$ (P. B. Duda et al., 2012; Yonce et al., 2019). However, the calibration stations had much smaller magnitude relative errors (Mean = 93.30%, Median = 0.49%) than those of the validation stations (Mean = 300.29%, Median = 24.29%, Table 5). Despite the limited performance in terms of relative error, 57.2% of the stations had KGE > -0.41 (Table 6), indicating that, for many stations, the model still improved upon using the mean observed value as a benchmark predictor.

3.5. Model performance by subbasin attribute

3.5.1. Discharge

Biases in discharge varied between geographical regions characterized by different catchment attributes (Fig. 7 & Table 7). In general, there was a slight negative correlation between discharge bias and latitude ($r = -0.28$, $p < 0.001$), and the model tended to overestimate discharges at lower latitudes while underestimating discharges at higher latitudes. Overall, the strongest correlations between discharge bias and catchment attributes were observed for the fraction of shallow soils in the Southeastern United Kingdom ($r = 0.53$, $p < 0.001$), deciduous forests in Southwestern Europe ($r = -0.51$, $p < 0.001$), perennial crops in Southwestern Europe ($r = 0.47$, $p < 0.001$), pastures in Northwest Scandinavia ($r = -0.40$, $p < 0.001$), and lakes in Southern Sweden ($r = 0.40$, $p < 0.001$).

Despite having similar dominant catchment characteristics, differing relationships were observed between the biases in discharge and the catchment characteristics for the North Central Europe, Southwestern Europe, and the Southeastern United Kingdom clusters. Although each of these three clusters were dominated by fine soils, agricultural and irrigated landuse, and spring and autumn crops (Fig. 7), the correlations and level of significance between the discharge biases and these catchment characteristics varied among the three clusters (Table 7). Furthermore, whereas the North Central Europe and Southwestern Europe clusters were the two clusters in the model domain with the greatest overestimation in discharge, discharges were slightly underestimated in the Southeastern United Kingdom cluster (Fig. 7). Overall, discharge biases were positively correlated with agriculture in Southwestern Europe ($r = 0.32$, $p < 0.01$) and with irrigated landuse in the Southeastern United Kingdom ($r = 0.39$, $p < 0.05$) and Southwestern Europe ($r = 0.32$, $p < 0.01$). Interestingly, discharge biases were correlated with fine soil positively in Northern Europe ($r = 0.22$, $p < 0.05$) but negatively in the Southeastern United Kingdom ($r = -0.39$, $p < 0.05$). In all three clusters, however, there were significant negative correlations between the discharge biases and the fraction of pasture landuse (Table 7). Throughout the full model domain, Southwestern Europe was the cluster with the highest overestimation in discharge; here the errors in discharge were also positively correlated with Mediterranean vegetation ($r = 0.28$, $p < 0.01$) and perennial crops ($r = 0.47$, $p < 0.001$), both of which were not prevalent in the other clusters.

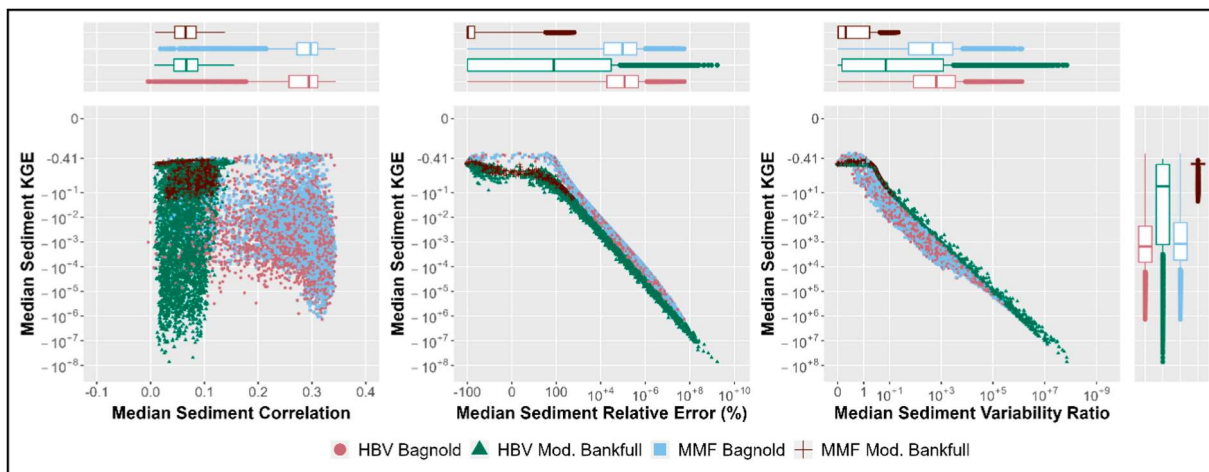


Fig. 7. Model performance for the 20,000 calibration ensemble members for the sediment calibration submodel. Each point represents the median model performance metric for a given candidate member based on all subbasins with observations within the submodel domain (319 calibration subbasins & 164 validation subbasins). Marginal boxplots are included to display the distribution of points along each axis, and the four combinations of HYPE's soil erosion models and stream sedimentation/resuspension routines are symbolized with different colors and marker shapes.

Table 7
Pearson correlation coefficients between scaled discharge relative error and catchment characteristics according to cluster. Significance levels are indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Dashes indicate clusters with no discharge observation stations with upstream areas containing any of the respective soil type or land use.

Catchment Characteristic		Cluster								
Category	Characteristic	All	Central Europe	North Central Europe	Northern Europe	Northwestern Scandinavia	South Central Europe	Southeastern United Kingdom	Southern Sweden	Southwestern Europe
General	Latitude	-0.28 ***	-0.11 **	0.02	-0.15 **	-0.24	-0.10	0.36 *	0.18	-0.40 ***
	Upstream Area (km ²)	0.06 *	0.05	0.01	0.03	0.21	0.08	-0.07	0.17	0.38 ***
	Mean Upstream Elevation (m)	0.03	-0.05	0.01	-0.15 **	0.05	0.20	0.01	0.13	0.19
	Mean Upstream Slope (%)	-0.04	-0.08 *	-0.02	-0.23 ***	0.11	0.10	0.02	0.22	-0.29 **
Soil Fraction	Shallow Soil	-0.01	0.01	0.17	-0.14 **	-0.06	0.09	0.53 ***	0.18	0.13
	Moraine Soil	-0.11 ***	-0.01	-0.03	0.00	0.00	0.15	—	0.03	-0.07
	Fine Soil	0.19 ***	0.16 ***	0.22 *	0.06	0.07	-0.05	-0.39 *	0.00	0.14
	Medium Soil	0.03	-0.09 **	-0.11	0.11 *	-0.11	-0.15	-0.08	-0.14	-0.15
Landuse Fraction	Glacier	0.05 *	0.01	—	-0.03	0.24 *	0.27 *	—	0.02	-0.13
	Lake	-0.06 *	0.04	-0.15	0.05	0.17	0.05	-0.06	0.40 ***	0.34 **
	Agriculture	0.25 ***	0.21 ***	0.11	0.14 **	-0.03	-0.13	0.25	-0.25 *	0.32 **
	Deciduous Forest	-0.06 *	-0.18 ***	0.26 *	0.07	—	-0.09	-0.21 ***	-0.06	-0.51 ***
	Open	-0.21 ***	-0.02	-0.19	-0.30 ***	-0.18	0.08	-0.08	0.01	-0.27 *
	Wetland	-0.15 ***	0.04	-0.12	-0.13 **	-0.02	-0.03	-0.13	0.02	0.36 ***
	Irrigated	0.18 **	0.12 **	0.08	0.07	0.20	0.22	0.39 *	-0.06	0.32 **
	Coniferous Forest	-0.06 *	-0.04	-0.17	0.14 **	0.05	-0.09	0.05	0.06	-0.07
	Northern Deciduous Forest	-0.14 ***	-0.04	—	-0.13 **	0.13	0.34 **	—	0.06	—
	Fores									
Crop Fraction	Mediterranean Pasture	0.22 ***	0.13 ***	—	—	—	0.11	—	—	0.28 **
	Mixed Forest	-0.01	-0.10 **	-0.25 *	0.11 *	-0.40 ***	-0.10	-0.33 *	-0.22	-0.37 ***
	No Crop	-0.08 **	-0.12 **	-0.10	-0.12 *	0.05	-0.16	0.06	0.28 *	-0.24 *
	Perennial Crop	-0.13 ***	0.09 **	-0.07	-0.23 ***	-0.14	0.21	0.04	0.38 ***	-0.19
	Spring & Autumn Crop	0.26 ***	0.19 ***	-0.14	0.04	0.20	0.05	-0.19	0.03	0.47 ***
		0.23 ***	0.19 ***	0.11	0.14 **	-0.03	-0.14	0.25	-0.25 *	0.15
	Other Crop	-0.10 ***	0.08 *	0.09	-0.18 ***	0.11	0.01	0.05	-0.02	0.20

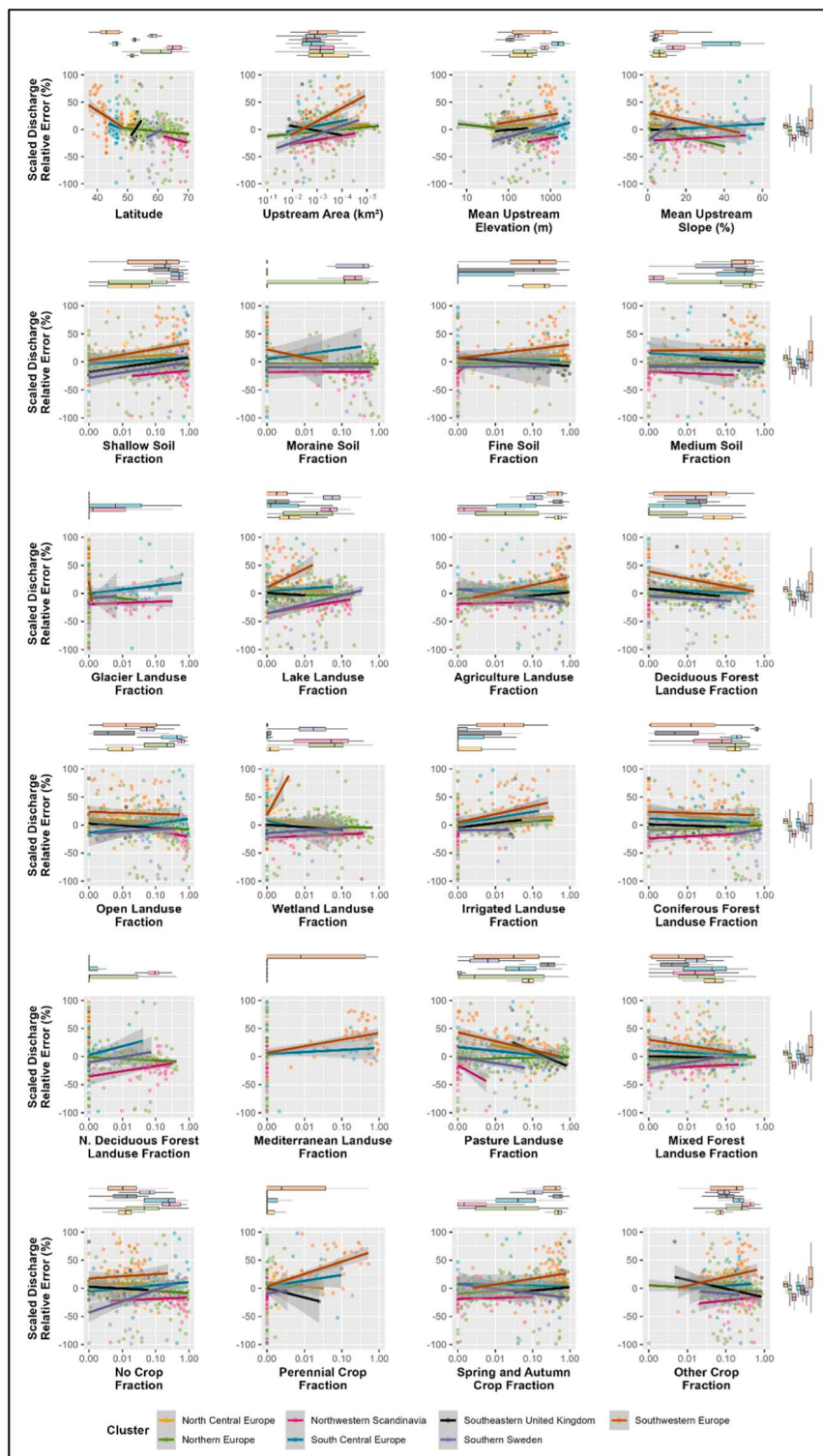


Fig. 8. Scaled relative error in discharge for the discharge calibration period (1991–2010) versus catchment characteristics for discharge observation stations within the clusters where precipitation correction factors were applied. Points represent individual observation stations, trendlines indicate the direction of relationships, and marginal boxplots show the distribution of values in each axis.

Table 8
Pearson correlation coefficients between scaled sediment relative error and catchment characteristics according to cluster. Significance levels are indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Dashes indicate clusters with no sediment observation stations with upstream areas containing any of the respective soil type or land use.

Catchment Characteristic		Cluster							
Category	Characteristic	All	Central Europe	North Central Europe	Northern Europe	South Central Europe	Southeastern United Kingdom	Southern Sweden	Southwestern Europe
General	Latitude	-0.11 **	-0.26 ***	-0.47 *	0.14	0.12	0.48	0.28	-0.25 **
	Upstream Area (km ²)	0.33 ***	0.36 ***	0.51 **	0.28 **	0.36	0.54 *	0.47 *	0.37 ***
	Mean Upstream Elevation (m)	0.40 ***	0.49 ***	0.50 *	0.54 ***	-0.30	0.58 *	0.40 *	0.49 ***
Soil Fraction	Mean Upstream Slope (%)	0.36 ***	0.43 ***	0.40	0.49 ***	-0.26	0.26	0.62 ***	0.46 ***
	Shallow Soil	0.27 ***	0.32 ***	0.32	0.19 *	0.14	0.61 **	-0.20	0.23 **
	Moraine Soil	-0.02	0.03	0.33	-0.18	-0.08	—	0.08	—
Landuse Fraction	Fine Soil	-0.30 ***	-0.29 ***	-0.17	-0.35 ***	-0.06	-0.43	—	-0.34 ***
	Medium Soil	0.04	0.04	0.30	-0.10	-0.06	0.32	-0.16	0.14
	Glacier	0.00	0.18 ***	—	—	-0.23	—	—	0.19 *
	Lake	0.07 *	0.00	0.16	0.29 **	0.04	0.27	0.31	-0.06
	Agriculture	-0.32 ***	-0.28 ***	-0.37	-0.28 **	-0.01	-0.15	-0.45 *	-0.59 ***
	Deciduous Forest	0.15 ***	0.11 **	0.23	0.02	0.16	-0.28	0.09	0.54 ***
	Open	0.30 ***	0.35 ***	0.32	0.54 ***	-0.34	0.29	0.50 **	0.43 ***
	Wetland	0.06	-0.08 *	0.58 **	0.43 ***	-0.16	0.20	-0.01	-0.12
	Irrigated	-0.09 **	-0.06	-0.05	-0.17	-0.22	-0.42	-0.26	-0.17
	Coniferous Forest	0.16 ***	0.25 ***	0.08	-0.04	0.01	0.07	0.24	0.24 **
	Northern Deciduous Forest	-0.05	-0.06	—	0.17	0.00	—	0.45 *	—
Crop Fraction	Mediterranean	0.06	0.04	—	—	0.14	—	—	0.06
	Pasture	-0.05	-0.03	0.27	-0.12	0.25	0.03	-0.25	-0.05
	Mixed Forest	0.28 ***	0.27 ***	0.48 *	0.17	0.43 *	-0.21	0.29	0.48 ***
	No Crop	0.06	0.04	0.11	0.10	-0.40	0.54 *	0.32	0.29 ***
	Perennial Crop	0.02	0.03	0.58 **	0.20 *	0.10	-0.09	0.09	-0.09
	Spring & Autumn Crop	-0.32 ***	-0.29 ***	-0.38	-0.28 **	-0.04	-0.15	-0.45 *	-0.57 ***
	Other Crop	0.17 ***	0.07	0.23	0.44 ***	-0.16	0.26	0.55 **	0.38 ***

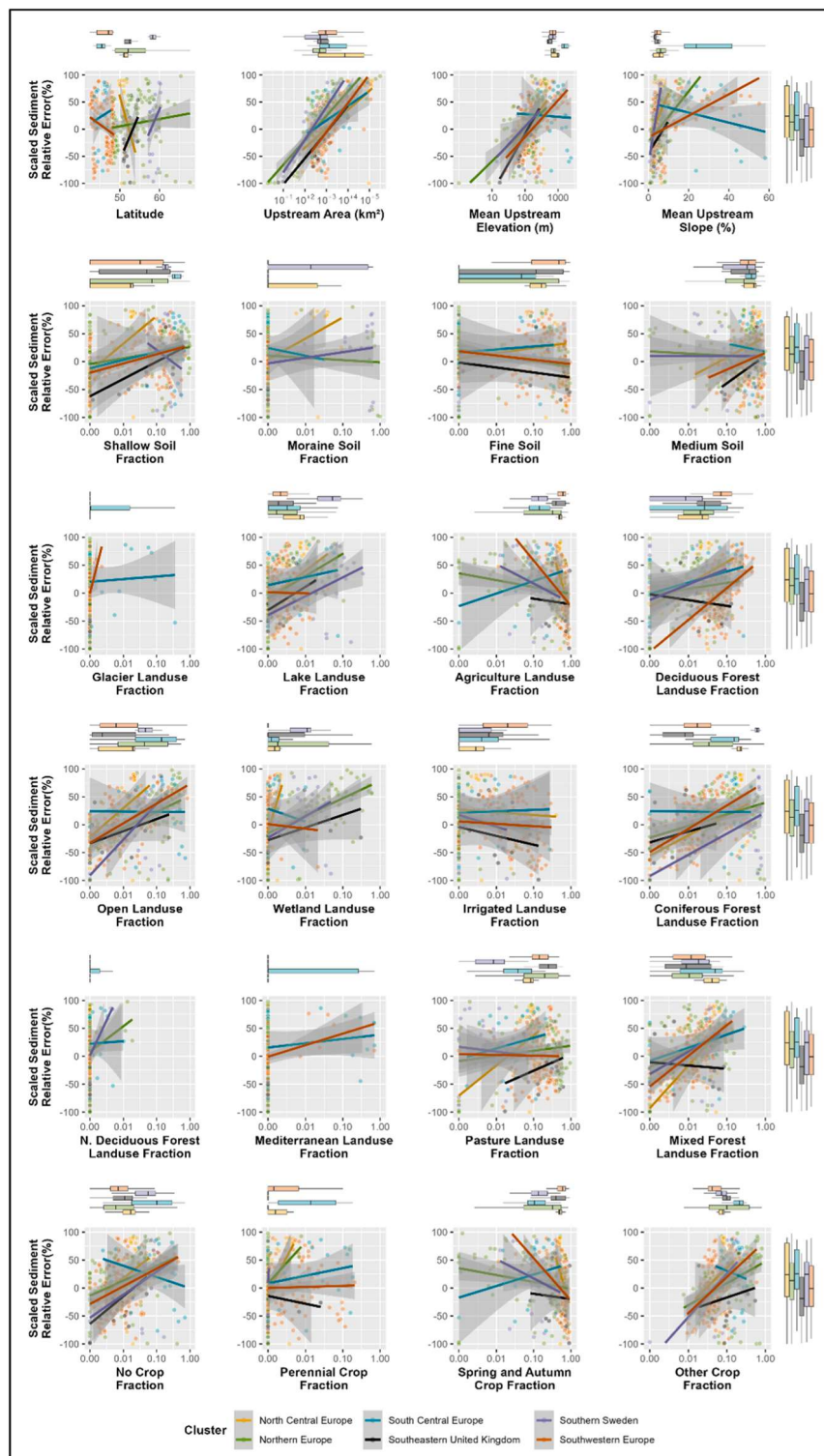


Fig. 9. Scaled relative error in sediment concentrations for the sediment calibration period (2001–2010) versus catchment characteristics for sediment observation stations within the clusters where precipitation correction factors were applied. Points represent individual observation stations, trendlines indicate the direction of relationships, and marginal boxplots show the distribution of values in each axis.

The Northern Europe, Northwestern Scandinavia, Southern Sweden, and South Central Europe also had similar catchment characteristics but different relationships between the biases in discharge and catchment characteristics. Across the model domain, the clusters with the greatest underestimation in discharge were the Northern Europe, Northwestern Scandinavia, and Southern Sweden clusters, which were characterized by high fractions of moraine soils and lake and wetland landuse (Fig. 7). In the Southern Sweden cluster, discharge bias was moderately correlated with lake fraction ($r = 0.40$, $p < 0.001$); however, discharge bias was not significantly correlated with the fraction of moraine soils in any of the three clusters, and only weakly correlated with wetlands in the Northern Europe cluster (Table 7). In comparison to the Northwestern Scandinavia cluster, the South Central Europe cluster was also characterized by high fractions of shallow soils, high fractions of glacier, open, and northern deciduous forest landuse, and no crops. Despite these similarities, however, the model tended to overestimate discharges in the South Central Europe cluster and underestimate discharges in the Northwestern Scandinavia cluster. In both of these clusters, there were slight positive correlations between discharge biases and glacier fraction ($p < 0.05$), but in the South Central Europe cluster, discharge bias was also weakly correlated with the fraction of northern deciduous forest ($r = 0.34$, $p < 0.01$).

3.5.2. Sediment

In general, sediment fluxes were overestimated across much of the model domain (Fig. 8), but there were stronger correlations between the sediment biases and catchment characteristics than between the discharge biases and catchment characteristics. Overall, sediment biases were significantly positively correlated with upstream area and mean upstream elevation in all but the South Central Europe cluster (Table 8). Significant positive correlations ($p < 0.001$) were also observed between sediment biases and mean upstream slope in the Northern Europe, Southern Sweden, and Southwestern Europe clusters. Among the individual clusters, the strongest correlations between sediment bias and catchment attributes were observed for the fraction of shallow soils in the Southeastern United Kingdom ($r = 0.61$, $p < 0.01$), agricultural landuse in Southwestern Europe ($r = -0.59$, $p < 0.001$), spring and autumn crops in Southwestern Europe ($r = -0.57$, $p < 0.001$), wetlands in North Central Europe ($r = 0.58$, $p < 0.01$), and perennial crops in North Central Europe ($r = 0.58$, $p < 0.01$). Fig. 9.

The greatest overestimation in sediment concentrations occurred in the South Central Europe cluster (Fig. 8), which often behaved differently from the other clusters. Unlike the other clusters, the South Central Europe cluster was distinguished by high fractions of perennial crops and large areas of glacier, northern deciduous forest, and Mediterranean vegetation landuse (Fig. 8). Overall, the South Central Europe cluster was the only cluster in which the biases in sediment were not significantly correlated to upstream area or mean upstream elevation (Table 8). In fact, the fraction of mixed forest landuse was the only catchment characteristic in the cluster for which there was a significant correlation with sediment biases ($r = 0.43$, $p < 0.05$).

The clusters with the smallest magnitude biases in sediment concentrations were the Northern Europe, Southwestern Europe, and Southeastern United Kingdom clusters. In each of these clusters, sediment biases were positively correlated with the fraction of shallow soils and negatively correlated with the fraction of fine soils (Table 8). Overall, however, correlations were stronger in the Northern Europe and Southwestern Europe clusters, and the sediment biases in both clusters were significantly positively correlated with open landuse and other crops while significantly negatively correlated with agricultural landuse and spring and autumn crops (Table 8); similar significant correlations were also observed between the sediment biases and these catchment characteristics in the Southern Sweden cluster.

Several other notable relationships were observed between sediment biases and catchment characteristics (Fig. 8 & Table 8). Significant positive correlations were observed between sediment biases and wetland and perennial crop fractions in the North Central Europe and Northern Europe clusters. In addition, there was a slight positive correlation between sediment bias and the fraction of lakes ($r = 0.29$, $p < 0.01$) in the Northern Europe cluster. In Southwestern Europe, there were significant positive correlations between sediment bias and the fraction of deciduous forests ($r = 0.54$, $p < 0.001$), coniferous forests ($r = 0.24$, $p < 0.01$), and mixed forests ($r = 0.48$, $p < 0.001$). Similar correlations with mixed forests were observed for both North Central Europe ($r = 0.48$, $p < 0.05$), and South Central Europe ($r = 0.43$, $p < 0.05$).

4. Discussion

4.1. Consequences of data availability for model calibration

The availability of representative calibration data was a critical factor in the transferability of the model parameters from the calibration domain to the full model domain. Overall, the model performance for discharge and evapotranspiration (Table 5) was similar between the calibration and validation stations. However, biases in discharge (Fig. 7) were larger in areas characterized by landuses with no or very few RGB basins: Lake, Glacier, Wetland, Irrigated (Agriculture & Pasture), Mediterranean Forests, and Northern Deciduous Forests. These landuse classes generally comprised only small portions of the model domain (Table 4), so there were few observation stations located in areas that were highly representative of these landuses. Although the selection of the sediment calibration stations was intended to provide a set of subbasins with similar distributions of catchment characteristics to those of the full model domain (Section 2.4.3), the lack of available sediment data in large parts of the model domain made it difficult to select a truly representative sample of subbasins. Consequently, the relative error between the simulated and observed sediment concentrations was much better for the calibration stations than for the validation stations (Tables 5 and 6).

Altogether, these results indicate that the overall calibration method — selecting representative basins, evaluating ensembles of parameter sets and model options, and ranking the candidates — allowed for good transferability of the model parameters from the calibration domain to the full model domain when calibration data was plentiful. However, the model parameters could not be

calibrated and transferred as successfully to areas where there was a lack of calibration data for the predominant landuses, as evident by the greater model biases in these parts of the model domain. Overall, this emphasizes the inherent challenge in large-scale modeling of finding sufficient calibration data to adequately describe the whole model domain. Fortunately, the increasing availability and resolution of alternative data sources, such as Earth Observations, could help to fill these data voids and improve the calibration of large-scale hydrological models.

4.2. Dominant factors and processes limiting model performance

Inspection of the model performance by subbasin attributes (Section 3.5) reveals the dominant factors and processes limiting model performance in various geographic regions. As highlighted in Section 4.1, model performance was limited by the lack of calibration data in areas with certain landuses. However, model structure and process representation also have important roles in model performance. Knowledge of the key processes limiting model performance in different geographic regions can not only help to prioritize future developments in the HYPE model code and target areas of improvement for further E-HYPE calibrations, but also assist other modelers in selecting which processes to focus calibration efforts on when developing other hydrological models in the European domain.

4.2.1. Discharge

4.2.1.1. Latitude and temperature. The slight negative correlation between discharge bias and subbasin latitude (Table 7) hints at several processes which may be significant for model performance across the full European domain. HYPE's Modified Jensen-Haise/McGuinness potential evapotranspiration routine (Section 2.2) uses subbasin latitude to estimate extraterrestrial solar radiation, and the calculated potential evapotranspiration is directly proportional to extraterrestrial solar radiation. Therefore, the relationship between discharge bias and latitude may be partially explained by the evapotranspiration processes across the model domain. In addition, both evapotranspiration and the partitioning of precipitation between rainfall and snowfall are dependent on the temperature forcing data. Thus, at higher latitudes, the model may also underestimate discharges because it underestimates evapotranspiration or overestimates the proportion of precipitation falling as snow and, consequently, generates less runoff.

4.2.1.2. Glaciers. Glacier area was a significant factor impacting the discharge biases in the South Central Europe and Northwestern Scandian clusters. These two clusters contained the highest fractions of glacier areas, and in both clusters, there were significant positive correlations ($p < 0.05$) between discharge bias and the fraction of glacier area (Table 7). Overall, this implies that the model does not store enough water in the glaciers. The snow melt routine used in E-HYPE4 is based on a temperature index in which snow melting occurs when the temperature is greater than a threshold temperature. However, this threshold temperature is uniform for all glacier areas, and this assumption may lead to an overall overestimation of glacier snow melt across the full model domain. HYPE includes alternative snow melt models in which snow melt can refreeze and for which snow melt is also affected by shortwave radiation, which can be derived according to subbasin latitude. Although these alternative model routines require additional calibration parameters, these routines could provide spatially varying glacier responses, which could potentially improve the model biases.

4.2.1.3. Irrigation. Irrigation greatly influenced the overestimation of discharge and evapotranspiration in the Southwestern Europe cluster. The southwestern portion of Europe is highly irrigated, and this cluster had the highest fractions of irrigated landuse and perennial crops (Fig. 7). Overall, the significant positive correlations between discharge biases and agriculture and irrigated landuses and perennial crops for this cluster (Table 7) suggest that irrigation may be the dominant process affecting model performance in this area. It is worth noting that the Southeastern United Kingdom cluster had the second highest fraction of irrigated areas and also had a significant positive correlation between discharge bias and the fraction of irrigated landuse. Although irrigation processes are included in E-HYPE4, the simulated amount of water extracted for irrigation is based on an estimated irrigation water demand (Section 2.2) rather than on actual observed/recorded data. Consequently, the model may underestimate the amount of water extracted for irrigation, leading to the overestimation in discharge. Furthermore, abstractions, water transfers, and other management processes are likely not fully described in the global datasets at the resolution used in E-HYPE4 and, therefore, may not be well represented in the model setup. The cumulative effect of these unmodeled processes may be reflected by the significant positive correlation between discharge bias and upstream catchment area in the Southwestern Europe cluster (Table 7). If detailed spatial data on abstractions were available across Europe, then model performance could be improved by explicitly including them in the model structure. Overall, this underscores the need for more detailed large-scale spatial datasets to improve model parameter values and model structure.

4.2.1.4. Groundwater. Groundwater processes were a key factor in the biases in discharge in the Southeastern United Kingdom. In this cluster, there was a significant positive correlation between discharge bias and the fraction of shallow soils (Table 7). Soils tend to be shallow in karst areas, which are prevalent in the southeastern portion of the United Kingdom. Although HYPE includes a regional aquifer module, this module was not used in the E-HYPE4 setup because it requires detailed local knowledge of the aquifer properties and groundwater connections. Thus, the positive correlation between bias and shallow soil fraction in this cluster is likely because the model is not capturing the transfer of water to the karst aquifers in the region. Hankin et al. (2019) developed a HYPE model setup for England and also attributed poorer model performance in southeastern England to the need for a groundwater module to link the surface catchments to the aquifers governing the baseflows across multiple catchments. Local refinements (e.g. addition of regional

aquifers) and calibration of the E-HYPE4 model setup in this area could help to improve model biases. However, this would require site-specific data and would introduce regionalization to the model setup.

4.2.1.5. Pastures. In many regions, biases in discharges were markedly linked to pasture areas. Overall, significant negative correlations were observed between discharge bias and pasture landuse in the North Central Europe, Northwestern Scandinavia, South-eastern United Kingdom, and Southwestern Europe clusters (Table 7). These findings suggest that the model underestimates the runoff produced in areas with pasture landuse. In HYPE's Modified Jensen-Haise/McGuinness potential evapotranspiration routine (Section 2.2), potential evapotranspiration is controlled by a landuse dependent coefficient. Consequently, the model may overestimate the evapotranspiration that occurs from the pasture landuse class, causing the model to produce less runoff from pasture areas.

The strongest correlation between discharge bias and the fraction of pasture landuse occurred in the Northwestern Scandinavia cluster, which was also the cluster with the greatest underestimation of discharges. The Northwestern Scandinavia cluster has a very small fraction of pasture landuse (Fig. 7), and discharge biases in the cluster were not significantly negatively correlated with any other catchment characteristic (Table 7). The underestimation of discharges in pasture areas may therefore have a tremendous impact on the water balance. However, an additional consideration is that the HydroGFD3.2 meteorological forcing data used by E-HYPE4 may underestimate the actual precipitation received in this region. Western Norway has a marine climate which receives a significant amount of precipitation, and comparison of the HydroGFD3 climatologies to other global precipitation products indicated that HydroGFD3 was generally drier in the area (Berg et al., 2021).

4.2.1.6. Lakes. The presence of lakes was an important contributor to the discharge biases in Southern Sweden. In this cluster, there was a significant positive correlation between discharge biases and the fraction of lake area (Table 7). In HYPE, a percentage of flow from the local streams is routed to the local lakes (Section 2.2). Sweden has approximately 100,000 lakes; thus, the correlation between discharge bias and lake area could indicate that the model is underestimating the amount of water that should be stored in the local lakes. An additional potential contributing factor to the correlation between discharge bias and lake area is that rivers and waterbodies are highly regulated in Sweden due to hydropower (Elenius and Lindström, 2022). However, detailed data on hydropower regulations are not publicly available, so these regulations cannot be modeled in detail in E-HYPE4. Thus, the positive correlation between discharge biases and lake area in the cluster may also be because the model releases more water from reservoirs than that which is actually released by the operators.

4.2.2. Sediments

4.2.2.1. Upstream area, elevation, and slope. Biases in sediment concentrations across the full European domain were positively correlated with upstream area, mean upstream elevation, and mean upstream slope (Table 8). The overestimation of sediment fluxes in large basins may be partially explained by the lack of sediment retention in the model. In HYPE's in-stream sedimentation and resuspension routines (Section 2.2), no particles are permanently removed from the simulation by sedimentation; thus, all eroded sediment particles can be eventually resuspended and redistributed over time. In addition, the erosion routines in HYPE do not account for gully/streambank erosion. Consequently, the model parameterization may overestimate the amount of soil erosion in order to compensate for the missing gully/streambank loads. However, the factors which drive gully/streambank erosion differ from those which drive soil erosion, and this may contribute to some of the poor sediment dynamics in the best global model setup (Table 5). The correlation between sediment biases and elevation likely reflects that the mean upstream elevation increases as drainage area increases. Steep slopes increase the amount and velocity of runoff and lead to greater soil erosion; thus, it is not surprising that there is a positive correlation between the sediment biases and mean upstream slope in the model. However, the HYPE soil erosion routines are based on the mean subbasin slope, and the model overestimation may indicate that the mean slope and HYPE's semi-distributed spatial representation are insufficient to capture all of the small-scale variations in slope across the model subbasins. Overall, these results stress the importance of process representation and model structure for model performance.

4.2.2.2. Agriculture. Agriculture was a significant factor influencing the biases in sediment concentrations in the Northern Europe, Southern Sweden, and Southwestern Europe clusters. These regions are highly agricultural, and in each of these clusters, sediment biases were significantly positively correlated with open landuse and other crops while significantly negatively correlated with the fraction of agriculture landuse and spring/autumn crops (Table 8). Altogether, this suggests that the model overestimates erosion from open areas while underestimating erosion from agricultural areas. In HYPE's MMF erosion model (Section 2.2), landuse affects soil erosion through ground and crop cover factors which reduce the amount of erosion due to rain and surface runoff. These factors vary seasonally according to specified crop calendars for each NUTS region; therefore, the biases in the open and agricultural areas may be related to the timing and magnitude of the cover factors which could lead to over/underestimation of the reductions in erosion due to vegetation. Manual adjustment of the crop calendars and cover factors for specific landuses could potentially reduce the sediment biases, but this would be tricky and time consuming. Overall, the three clusters had the smallest magnitude biases in sediment concentrations (Fig. 8). However, this may simply reflect that the underestimation of the sediment fluxes from the agricultural areas was large enough to offset the general overestimation of sediments across the full model domain. In evidence of this, among the three clusters, the Southwestern Europe cluster had simultaneously the largest fraction of agriculture landuse and the smallest median sediment bias whereas the Southern Sweden cluster had simultaneously the smallest fraction of agriculture landuse and the greatest median sediment bias (Fig. 8).

4.2.2.3. Forests. Throughout the model domain, and in Southwestern Europe especially, biases in sediment concentrations were strongly associated with the fraction of forest areas. In general, the model appears to overestimate the amount of erosion which occurs in forest areas, as evidenced by the significant positive correlations between sediment biases and the fraction of deciduous, coniferous, and mixed forests for the full domain (Table 8). Among the individual clusters, the correlations between sediment bias and forest area were strongest in the Southwestern Europe cluster, and significant positive correlations were also observed for deciduous, coniferous, and mixed forests. Likewise, sediment biases were significantly positively correlated with northern deciduous forest in Southern Sweden and with mixed forest in North Central Europe and South Central Europe. Altogether, these results indicate that the model overestimates the amount of erosion which occurs in forest areas. As previously described, these biases could occur if the MMF erosion model's crop calendars and ground/crop cover factors underestimate the impact of forest vegetation on reducing the erosion due to rainfall and surface runoff. Forest areas are considered permanent vegetation in HYPE, and constant (i.e. not seasonally varying) ground/crop factors are used throughout the simulation for these landuse classes. Thus, the erosion model does not account for intra-annual changes in the interception capacity of the forest land cover, and the model parameterization may be biased to compensate for this. In total, there were 315 separate crop calendars defined for the forest landuse classes in E-HYPE4, so tuning of the MMF cover factors would be difficult. Altogether, this emphasizes the importance of the static (i.e. non-calibration) model parameterization and catchment descriptions on the model performance.

4.2.2.4. Wetlands and lakes. In the North Central Europe and Northern Europe clusters, the biases in sediment concentrations were related to the presence of wetlands and lakes. For both clusters, sediment biases were significantly positively correlated with the fraction of wetland area (Table 8). Although wetlands are effective sediment traps, HYPE does not explicitly simulate wetlands. Thus, the lack of sediment retention within the model's wetland landuse class likely contributes to the overestimation of sediments in these clusters. In the Northern Europe cluster, sediment biases were also significantly positively correlated with the fraction of lake area. There are many small lakes in Northern Europe, and this suggests that the model underestimates the trapping of sediments within lakes in this region. In E-HYPE4, the lake settling parameters are uniform across the whole model domain. HYPE includes the ability to specify the sedimentation velocity for specific lakes, and this could potentially provide improvements in the model sediment biases. However, this would require either site-specific data or manual tuning of the model on a lake by lake basis. Alternatively, a new HYPE routine could be developed to calculate lake settling according to e.g. various lake properties.

4.3. Significance of sediment model routine selection

4.3.1. Erosion model

Altogether, the results from this study demonstrate that increased model routine complexity does not necessarily correspond to increased model performance. The index-based HBV erosion model was developed as a less complex alternative to the process-based MMF routine and was intended for modeling large domains and data scarce regions where the detailed information on soils, crops, and agricultural practices required by the MMF model is not available. However, the results from this study showed that the HBV model performed equally well as the MMF model and indicate that the simplifications in the HBV model do not imply a commensurate limitation in model performance.

These results also highlight the limitations in performance that can be imposed by the model structure when using process-based routines. While the MMF erosion model has been used to assess soil erosion at large scales and daily time steps (Eekhout et al., 2018; Jain and Ramsankaran, 2018; Parente et al., 2022; Shrestha and Jetten, 2018), the model was originally developed to predict annual soil losses from field-sized areas on hillslopes. In HYPE's implementation of the MMF model, mobilized sediments are calculated individually for each of the model SLC classes and aggregated to compute the total flux of eroded sediment from each subbasin. However, at the temporal and spatial resolution used in E-HYPE4 (daily time step and median subbasin area of 170 km²), the MMF routine may be unable to capture all of the small-scale variations in the catchment characteristics (e.g. slopes, soils, landuse, etc.) and erosion processes present in each subbasin. Thus, in E-HYPE4, the additional complexity of the process-based MMF model may not provide significant benefit in comparison to the index-based structure of the HBV model, and this may explain why the two routines provided similar performance.

In regards to the model performance for sediments, the choice of erosion model appeared to be less important than the choice of in-stream sedimentation/resuspension method for the E-HYPE4 model setup. This was especially apparent when analyzing the correlations between the simulated and observed sediment concentrations, where the differences in correlations were almost entirely due to the choice of sedimentation/resuspension routine. Because the choice of erosion model did not seem to affect the correlation values, this could potentially indicate that the two erosion models are not capturing the dynamics in the overland processes and that the simulated sediment dynamics are coming from the in-stream processes. It is important to note that HYPE's MMF and HBV erosion models do not explicitly account for gully or streambank erosion, which can contribute a large percentage of the total sediment export from a basin (Gellis and Noe, 2013; Palmer et al., 2014; Poesen et al., 2003; Sidle et al., 2019; Theuring et al., 2013; Walling, 2005; Walling et al., 2008). Consequently, in E-HYPE4, the model may rely on the in-stream sedimentation and resuspension processes to compensate for the unmodeled gully/streambank sediment sources, and this may partially explain the model's higher dependence on the stream sedimentation/resuspension routine than on the selected erosion model. Therefore, the relative importance of the selection of erosion model vs. stream sedimentation/resuspension method may vary for other HYPE model setups depending on model/site specific factors such as subbasin size, the dominant sources of eroded sediments in the domain, and the significance of overland vs. in-stream sediment processes.

4.3.2. In-stream sedimentation/resuspension method

Analysis of the sediment calibration ensemble results (Section 3.3) indicated that different KGE components were optimized by the two in-stream sedimentation/resuspension methods. For the ensemble members using the Simplified Bagnold Equation routine, KGE was optimized by maximizing the correlation and minimizing the magnitude of the variability ratio. In contrast, for the members using the Modified Bankfull Flow routine, KGE was optimized by minimizing the magnitude of the relative errors. Overall, however, both sedimentation/resuspension routines were able to provide similar maximum KGE values. Therefore, these results suggest that it may be advantageous to select different routines for different use cases. The Bankfull Flow routine may be a better choice for studies of sediment loads to coastal areas, for example, where minimizing the bias in simulated sediment concentrations may be of greater importance than matching the daily sediment dynamics. Conversely, for studies where matching the sediment dynamics is more important, the Bagnold Equation may provide better results.

The differences in the model performance components between the two sedimentation/resuspension methods may be due to differences in their process descriptions. Compared to the Bankfull Flow routine, the Bagnold routine provided much higher correlation values (Fig. 6). This difference could be due to the contrasting processes which control the partitioning between sedimentation/resuspension in the two routines; in the Bagnold routine, sedimentation/resuspension is dependent on the simulated sediment concentration, whereas in the Bankfull Flow routine, sedimentation/resuspension is controlled by the simulated discharge. Furthermore, whereas the Bankfull Flow routine does not account for streambank erosion, HYPE's implementation of the Bagnold routine allows the amount of resuspended sediment to be increased up to the maximum transport capacity in order to represent potential bank/channel erosion (Section 2.2). Thus, the higher correlations with the Bagnold routine may also occur because the routine is able to account for some of the contributions from these additional sediment sources. However, the additional sediment added to the system by the Bagnold routine may explain why the routine typically resulted in higher relative errors than those produced by the Bankfull Flow routine (Fig. 6). HYPE doesn't account for sediment retention in basins, and all suspended sediment is eventually resuspended and transported downstream. Therefore, the development of a model routine to permanently trap/remove a fraction of the mobilized sediments in each subbasin could potentially help to reduce the overestimation of sediment concentrations that occurs with the Bagnold routine.

4.4. Impacts of regionalization on model performance

Increasing the regionalization of large-scale hydrological models can compensate for the limitations in model performance imposed by simulating the many critical processes for different regions (Section 4.2) within a single model setup. In Fig. 2, comparing the maximum achievable performance with the number of candidate model setups reveals how model performance was affected by the potential degree of regionalization. This assumes, however, that the model performance was independent for each station and that, for example, in the cases of upstream and downstream catchments along the same river, separate parameter sets could be used to calculate the outputs for each location. Throughout the calibration process, the maximum achievable performance was greatest when the greatest number of candidate parameter sets were included in a calibration step, and the model could be the most regionalized (Fig. 2 Steps 1, 3, & 9). Inclusion of regional precipitation corrections during the Water Balance Check phase of calibration also significantly improved model performance (Fig. 2, Steps 3 & 4). However, the maximum achievable performance decreased only slightly when going from the discharge and sediment calibration submodels with 20,000 candidate members (Fig. 2, Steps 1, 3, & 9) to the full domain model with the top performing candidates (Fig. 2, Steps 2, 4, & 10). Thus, because the performance from the full 20,000-member ensemble was emulated with $\leq \approx 3\%$ of the ensemble members, the performance of a highly regionalized model could be achieved with a much less regionalized model. The largest drops in achievable performance occurred when going from an ensemble of candidate parameter sets to using a single parameter set for the full model domain (Fig. 2, Steps 5 & 11), which indicates that there are some processes and/or catchment descriptors that cannot be fully captured by a single model setup. This highlights the tradeoff between regionalization and model performance for large-scale models; increased regionalization provides better achievable performance but can increase the uncertainty of how this regionalization will affect simulations under scenarios with different climate condition and catchment developments.

A compromise between regionalization and model performance could potentially be obtained by adapting the calibration process to select an optimal ensemble of only a few (e.g. 5–10) model setups, which could provide some of the improved model performance provided by regionalization but reduce some of the uncertainty associated with highly regionalized models. When calibrating hydrological models, the best fitting parameter values can compensate for structural errors in the model setup (Gupta and Govindaraju, 2019). Thus, regionalized models may help to improve model performance by applying different parameter values over different parts of the model domain, but the parametric and structural uncertainties associated with the model setup remain. In contrast, using an ensemble of model setups with different parameter sets can reduce the parametric uncertainties associated with equifinality in the model calibration. Furthermore, multi-parameter ensembles have been found to more effectively improve model performance compared to single-model, single-parameter, and multi-model approaches (Shin et al., 2023).

4.5. Implications of multi-objective calibration for equifinality

Analysis of the maximum achievable performance throughout the various steps of the calibration process (Section 3.1) provides some insight into the success of the multi-objective ensemble calibration scheme in reducing equifinality in the calibration. During the discharge calibration (Fig. 2, Steps 3–5), less substantial reductions in achievable model performance were observed for evapotranspiration than for discharge and sediments during selection of the top performing parameter sets (Fig. 2, Step 4) and identification

of the single parameter set providing the optimal full domain model performance (Fig. 2, Step 5). This indicates that there were many candidate ensemble members that could satisfactorily model the long-term averages in potential evapotranspiration, but fewer that could also produce acceptable simulations for discharge and sediments. The difference in the magnitudes of the biases in discharge and evapotranspiration may reflect their relative proportions in the water balance; discharge may only be a small fraction of rainfall, so small changes in the model parameters may have a larger impact on discharge than on evapotranspiration. Although calibration of the model for sediments (Fig. 2, Steps 9–11), did not meaningfully affect the achievable model performance for discharge and evapotranspiration, achievable model performance for sediments was significantly limited by the selection of the optimal parameter set during the discharge calibration (Fig. 2, Step 5). Throughout the sediment calibration, there were notable improvements in sediment KGE, but less substantial improvements in sediment relative error. This suggests that the sediment calibration could not compensate for biases in the discharge calibration. Overall, these findings suggest that, although it was useful as a check of the water balance, inclusion of evapotranspiration as a calibration objective was of limited use in reducing equifinality in the calibration because acceptable model performance for evapotranspiration could be achieved by a large fraction of the ensemble members. It is worth noting, however, that the evapotranspiration observations used in the model calibration were long-term average values. Consequently, using e.g. daily evapotranspiration observations rather than long-term average observations could provide more value in reducing equifinality because the dynamics in evapotranspiration could also be assessed. Compared to long-term average evapotranspiration, the inclusion of sediments as a calibration objective provided greater value in reducing equifinality during the discharge calibration because the model performance for sediments was dependent on the discharge calibration while the model performance for discharge was independent of the sediment calibration.

5. Conclusions

Calibrating a large-scale hydrological model will always be a compromise among factors such as data availability, complexity of process descriptions and model structure, regionalization, and model performance. The development and calibration of the E-HYPE4 model provided some insights into these tradeoffs, and the primary conclusions drawn from this study are summarized as follows:

- The calibration scheme provided good transferability of the model parameters from the calibration domain to the full model domain for areas dominated by soil and landuse classes for which calibration data was plentiful. However, model biases were larger in areas which were characterized by landuse classes for which there was little to no available calibration data, which emphasizes the inherent challenge in large-scale modeling of acquiring sufficient calibration data to describe the entire model domain.
- Analysis of model performance by subbasin attributes revealed the dominant processes affecting model biases in discharge and sediments for various geographic regions and prevailing catchment characteristics.
- The simpler, index-based HBV soil erosion model performed equally well as the more complex, process-based MMF model. For the in-stream sedimentation/resuspension routines, better correlations between the simulated and observed sediment concentrations were achieved with the Simplified Bagnold Equation while smaller biases were provided by the Modified Bankfull Flow method.
- Model performance was best when the degree of regionalization was the highest, and the largest drop in maximum achievable model performance occurred when going from an ensemble of parameter sets to using a single parameter set for the full model domain. However, much of the gains in performance with a highly regionalized model could be achieved with a much less regionalized model.
- The inclusion of sediments as a calibration objective provided more value than that of evapotranspiration in regards to reducing equifinality in the model calibration.

It is important to acknowledge the limitations of this study. First, the analysis in this study is based on a single modeling software (HYPE) and a single model setup (E-HYPE4). Consequently, some of the limitations in model performance identified in this study could potentially be addressed with other modeling software or model setups with, for example, different process representations or land descriptions (e.g. soil and landuse classes). In addition, the model was evaluated with only one meteorological forcing product, HydroGFD, so some of the simulation errors could also be derived from uncertainty and/or errors in the input forcing data. One noteworthy limitation of HYPE is that the landuse description is static throughout the simulation period. Thus, impacts due to changes in landuse are not reflected in the model simulations, and some of the model errors may result from situations in which the modeled upstream landuse for an observation station is not representative of the actual upstream landuse when the observations were made. Furthermore, HYPE's sediment routines also do not explicitly account for gully/streambank erosion or sediment retention in sub-basins, and these unmodeled processes can have important implications for sediment fluxes. Finally, as a large-scale model, E-HYPE4's spatial representation may not be conducive to taking full advantage of certain process-based routines such as HYPE's MMF erosion model.

CRedit authorship contribution statement

Conrad Brendel: Conceptualization, Methodology, Formal analysis, Writing – Original Draft, Visualization. **René Capell:** Conceptualization, Methodology, Writing – Review & Editing. **Alena Bartosova:** Conceptualization, Writing – Review & Editing, Supervision, Funding acquisition.

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.

Data Availability

Data will be made available on request.

Acknowledgements

The investigation was performed at the SMHI Hydrological Research unit, where much work is done jointly, and we would like to acknowledge all of our co-workers. This project was part of AXIS, an ERA-NET initiated by JPI Climate, and funded by FFG Austria, BMBWF Austria, BMBF Germany, NWO Netherlands, RCN Norway, and FORMAS Sweden with co-funding by the European Union [grant number 776608]. The HYPE model code is open source and can be retrieved with documentation at <http://hypeweb.smhi.se>. Special thanks to Iron Maiden always being a great source of motivation and inspiration \m/.

References

- Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56.
- Arsenault, R., Breton-Dufour, M., Poulin, A., Dallaire, G., Romero-Lopez, R., 2019. Streamflow prediction in ungauged basins: analysis of regionalization methods in a hydrologically heterogeneous region of Mexico. *Hydrol. Sci. J.* 64, 1297–1311. <https://doi.org/10.1080/02626667.2019.1639716>.
- Ashraf Vaghefi, S., Iravani, M., Sauchyn, D., Andreichuk, Y., Goss, G., Faramarzi, M., 2019. Regionalization and parameterization of a hydrologic model significantly affect the cascade of uncertainty in climate-impact projections. *Clim. Dyn.* 53, 2861–2886. <https://doi.org/10.1007/s00382-019-04664-w>.
- Assunção, R.M., Neves, M.C., Câmara, G., Da Costa Freitas, C., 2006. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *Int. J. Geogr. Inf. Sci.* 20, 797–811. <https://doi.org/10.1080/13658810600665111>.
- Bartosova, A., Arheimer, B., de Lavenne, A., Capell, R., Strömqvist, J., 2021. Large-Scale Hydrological and Sediment Modeling in Nested Domains under Current and Changing Climate. *J. Hydrol. Eng.* 26, 05021009. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0002078](https://doi.org/10.1061/(ASCE)HE.1943-5584.0002078).
- Berg, P., Almén, F., Bozhinova, D., 2021. HydroGFD3.0 (Hydrological Global Forcing Data): a 25 km global precipitation and temperature data set updated in near-real time. *Earth Syst. Sci. Data* 13, 1531–1545. <https://doi.org/10.5194/essd-13-1531-2021>.
- Blöschl, G., 2013. *Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales*. Cambridge University Press, Cambridge.
- Brendel, C.E., Dymond, R.L., Aguilar, M.F., 2021. Modeling storm sewer networks and urban flooding in Roanoke, Virginia, with SWMM and GSSHA. *J. Hydrol. Eng.* 26, 05020044. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0002021](https://doi.org/10.1061/(ASCE)HE.1943-5584.0002021).
- Budhathoki, S., Rokaya, P., Lindenschmidt, K.-E., Davison, B., 2020. A multi-objective calibration approach using in-situ soil moisture data for improved hydrological simulation of the Prairies. *Hydrol. Sci. J.* 65, 638–649. <https://doi.org/10.1080/02626667.2020.1715982>.
- Butts, M.B., Payne, J.T., Kristensen, M., Madsen, H., 2004. An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation (The Distributed Model Intercomparison Project (DMIP)). *J. Hydrol.* 298, 242–266. <https://doi.org/10.1016/j.jhydrol.2004.03.042>.
- Casado-Rodríguez, J., del Jesus, M., 2022. Hydrograph separation for tackling equifinality in conceptual hydrological models. *J. Hydrol.* 610, 127816. <https://doi.org/10.1016/j.jhydrol.2022.127816>.
- Chapman, W.L., Welch, W.J., Bowman, K.P., Sacks, J., Walsh, J.E., 1994. Arctic sea ice variability: Model sensitivities and a multidecadal simulation. *J. Geophys. Res.* 99, 919. <https://doi.org/10.1029/93JC02564>.
- , 2012dataset EEA, 2012. The European Pollutant Release and Transfer Register (E-PRTR) v3.3. https://cdr.eionet.europa.eu/help/eptr_lcp/Documents/EPTR-LCP_datamodel_v3.3.pdf.
- de Lavenne, A., Lindström, G., Strömqvist, J., Pers, C., Bartosova, A., Arheimer, B., 2022. Evaluation of overland flow modelling hypotheses with a multi-objective calibration using discharge and sediment data. *Hydrol. Process.* 36, e14767. <https://doi.org/10.1002/hyp.14767>.
- Di Marco, N., Avesani, D., Righetti, M., Zaramella, M., Majone, B., Borga, M., 2021. Reducing hydrological modelling uncertainty by using MODIS snow cover data and a topography-based distribution function snowmelt model. *J. Hydrol.* 599, 126020. <https://doi.org/10.1016/j.jhydrol.2021.126020>.
- Donnelly, C., Andersson, J.C.M., Arheimer, B., 2016. Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. *Hydrol. Sci. J.* 61, 255–273. <https://doi.org/10.1080/02626667.2015.1027710>.
- Duda, P.B., Hummel, P.R., Donigan Jr., A.S., Imhoff, J.C., 2012. BASINS/HSPF: Model Use, Calibration, and Validation. *Trans. ASABE* 55, 1523–1547. <https://doi.org/10.13031/2013.42261>.
- Eekhout, J.P.C., Terink, W., de Vente, J., 2018. Assessing the large-scale impacts of environmental change using a coupled hydrology and soil erosion model. *Earth Surf. Dyn.* 6, 687–703. <https://doi.org/10.5194/esurf-6-687-2018>.
- Eini, M.R., Massari, C., Piniewski, M., 2023. Satellite-based soil moisture enhances the reliability of agro-hydrological modeling in large transboundary river basins. *Sci. Total Environ.* 873, 162396. <https://doi.org/10.1016/j.scitotenv.2023.162396>.
- Elenius, M.T., Lindström, G., 2022. Introduced flow variability and its propagation downstream of hydropower stations in Sweden. *Hydrol. Res.* 53, 1321–1339. <https://doi.org/10.2166/nh.2022.138>.
- [dataset] ESA, 2016. CCI Land Cover dataset (v 1.6.1). (<http://maps.elie.ucl.ac.be/CCI/viewer/download.php>).
- [dataset] European Union, Copernicus Land Monitoring Service, 2019. EU-Hydro – River Network Database, Version 1.3. (<https://land.copernicus.eu/imagery-in-situ/eu-hydro/eu-hydro-river-network-database>).
- [dataset] European Union, Copernicus Land Monitoring Service, 2020. Corine Land Cover (CLC) 2006, Version 2020_20u1. (<https://land.copernicus.eu/pan-european/corine-land-cover/clc-2006>).
- [dataset] Eurostat, 2022. Statistical regions in the European Union and partner countries. NUTS and statistical regions 2021: 2022 edition. Eurostat 188. (<https://ec.europa.eu/eurostat/web/nuts/background>).
- [dataset] FAO, 2003. The Digital Soil Map of the World Version 3.6. (<https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/faunesco-soil-map-of-the-world/en/>).
- Fernandez-Palomino, C.A., Hattermann, F.F., Krysanova, V., Vega-Jácume, F., Bronstert, A., 2021. Towards a more consistent eco-hydrological modelling through multi-objective calibration: a case study in the Andean Vilcanota River basin, Peru. *Hydrol. Sci. J.* 66, 59–74. <https://doi.org/10.1080/02626667.2020.1846740>.
- Finger, D., Vis, M., Huss, M., Seibert, J., 2015. The value of multiple data set calibration versus model complexity for improving the performance of hydrological models in mountain catchments. *Water Resour. Res.* 51, 1939–1958. <https://doi.org/10.1002/2014WR015712>.
- Gellis, A.C., Noe, G.B., 2013. Sediment source analysis in the Lingonore Creek watershed, Maryland, USA, using the sediment fingerprinting approach: 2008 to 2010. *J. Soils Sediment.* 13, 1735–1753. <https://doi.org/10.1007/s11368-013-0771-6>.

- Gentilucci, M., Bufalini, M., D'Aprile, F., Materazzi, M., Pambianchi, G., 2021. Comparison of Data from Rain Gauges and the IMERG Product to Analyse Precipitation in Mountain Areas of Central Italy. *ISPRS Int. J. Geo-Inf.* 10, 795. <https://doi.org/10.3390/ijgi10120795>.
- Ghaffar, S., Jomaa, S., Meon, G., Rode, M., 2021. Spatial validation of a semi-distributed hydrological nutrient transport model. *J. Hydrol.* 593, 125818 <https://doi.org/10.1016/j.jhydrol.2020.125818>.
- Gupta, A., Govindaraju, R.S., 2019. Propagation of structural uncertainty in watershed hydrologic models. *J. Hydrol.* 575, 66–81. <https://doi.org/10.1016/j.jhydrol.2019.05.026>.
- 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, 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>.
- Hankin, B., Strömquist, J., Burgess, C., Pers, C., Bielby, S., Revilla-Romero, B., Pope, L., 2019. A New National Water Quality Model to Evaluate the Effectiveness of Catchment Management Measures in England. *Water* 11, 1612. <https://doi.org/10.3390/w11081612>.
- Hartmann, J., Lauerwald, R., Moosdorf, N., 2014. A Brief Overview of the GLObal River Chemistry Database, GLORICH. *Geochemistry of the Earth's surface GES-10 Paris France, 18-23 August, 2014 Procedia Earth Planet. Sci.* 10, 23–27. <https://doi.org/10.1016/j.proeps.2014.08.005>.
- [dataset]Hartmann, J., Lauerwald, R., Moosdorf, N., 2019. GLORICH - Global river chemistry database. (<https://doi.org/10.1594/PANGAEA.902360>).
- Her, Y., Seong, C., 2018. Responses of hydrological model equifinality, uncertainty, and performance to multi-objective parameter calibration. *J. Hydroinformatics* 20, 864–885. <https://doi.org/10.2166/hydro.2018.108>.
- Her, Y., Yoo, S.-H., Cho, J., Hwang, S., Jeong, J., Seong, C., 2019. Uncertainty in hydrological analysis of climate change: multi-parameter vs. multi-GCM ensemble predictions. *Sci. Rep.* 9, 4974. <https://doi.org/10.1038/s41598-019-41334-7>.
- Hundecha, Y., Arheimer, B., Donnelly, C., Pechlivanidis, I., 2016. A regional parameter estimation scheme for a pan-European multi-basin model. *J. Hydrol. Reg. Stud.* 6, 90–111. <https://doi.org/10.1016/j.ejrh.2016.04.002>.
- Jain, P., Ramsankaran, R., 2018. GIS-based modelling of soil erosion processes using the modified-MMF (MMMF) model in a large watershed having vast agro-climatological differences. *Earth Surf. Process. Landf.* 43, 2064–2076.
- Jones, D.R., Schonlau, M., Welch, W.J., 1998. Efficient Global Optimization of Expensive Black-Box Functions. *J. Glob. Optim.* 13, 455–492. <https://doi.org/10.1023/A:1008306431147>.
- Karlsson, I.B., Sonnenborg, T.O., Refsgaard, J.C., Trolle, D., Børgesen, C.D., Olesen, J.E., Jeppesen, E., Jensen, K.H., 2016. Combined effects of climate models, hydrological model structures and land use scenarios on hydrological impacts of climate change. *J. Hydrol.* 535, 301–317. <https://doi.org/10.1016/j.jhydrol.2016.01.069>.
- Khatami, S., Peel, M.C., Peterson, T.J., Western, A.W., 2019. Equifinality and Flux Mapping: A New Approach to Model Evaluation and Process Representation Under Uncertainty. *Water Resour. Res.* 55, 8922–8941. <https://doi.org/10.1029/2018WR023750>.
- Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrol. Earth Syst. Sci.* 23, 4323–4331. <https://doi.org/10.5194/hess-23-4323-2019>.
- Kotteck, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World Map of the Köppen–Geiger climate classification updated. *Meteorol. Z.* 259–263. <https://doi.org/10.1127/0941-2948/2006/0130>.
- Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., Kundzewicz, Z.W., 2018. How the performance of hydrological models relates to credibility of projections under climate change. *Hydrol. Sci. J.* 63, 696–720. <https://doi.org/10.1080/02626667.2018.1446214>.
- Liu, X., Yang, K., Ferreira, V.G., Bai, P., 2022. Hydrologic Model Calibration With Remote Sensing Data Products in Global Large Basins. *Water Resour. Res.* 58, e2022WR032929 <https://doi.org/10.1029/2022WR032929>.
- Loepky, J.L., Sacks, J., Welch, W.J., 2009. Choosing the Sample Size of a Computer Experiment: A Practical Guide. *Technometrics* 51, 366–376. <https://doi.org/10.1198/TECH.2009.08040>.
- Mizukami, N., Clark, M.P., Newman, A.J., Wood, A.W., Gutmann, E.D., Nijssen, B., Rakovec, O., Samaniego, L., 2017. Towards seamless large-domain parameter estimation for hydrologic models. *Water Resour. Res.* 53, 8020–8040. <https://doi.org/10.1002/2017WR020401>.
- Morgan, R.P.C., Morgan, D., Finney, H.J., 1984. A predictive model for assessment of erosion risk. *J. Agric. Eng. Res* 30, 245–253. [https://doi.org/10.1016/S0021-8634\(84\)80025-6](https://doi.org/10.1016/S0021-8634(84)80025-6).
- Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* 50, 885–900.
- Mu, Q., Zhao, M., Running, S.W., 2011. Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sens. Environ.* 115, 1781–1800. <https://doi.org/10.1016/j.rse.2011.02.019>.
- Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R., 2011. Soil and water assessment tool theoretical documentation version 2009. Texas Water Resources Institute, (<https://swat.tamu.edu/media/99192/swat2009-theory.pdf>).
- Nemri, S., Kinnard, C., 2020. Comparing calibration strategies of a conceptual snow hydrology model and their impact on model performance and parameter identifiability. *J. Hydrol.* 582, 124474 <https://doi.org/10.1016/j.jhydrol.2019.124474>.
- Orth, R., Staudinger, M., Seneviratne, S.I., Seibert, J., Zappa, M., 2015. Does model performance improve with complexity? A case study with three hydrological models. *J. Hydrol.* 523, 147–159. <https://doi.org/10.1016/j.jhydrol.2015.01.044>.
- Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., Loumagne, C., 2005. Which potential evapotranspiration input for a lumped rainfall–runoff model? Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. *J. Hydrol.* 303, 290–306. <https://doi.org/10.1016/j.jhydrol.2004.08.026>.
- Palmer, J.A., Schilling, K.E., Isenhardt, T.M., Schultz, R.C., Tomer, M.D., 2014. Streambank erosion rates and loads within a single watershed: Bridging the gap between temporal and spatial scales. *Geomorphology* 209, 66–78. <https://doi.org/10.1016/j.geomorph.2013.11.027>.
- Panagos, P., 2006. The European soil database. *GEO Connex.* 5, 32–33. <https://esdac.jrc.ec.europa.eu/>.
- Parajka, J., Viglione, A., Rogger, M., Salinas, J.L., Sivapalan, M., Blöschl, G., 2013. Comparative assessment of predictions in ungauged basins - Part 1: Runoff-hydrograph studies. *Hydrol. Earth Syst. Sci.* 17, 1783–1795. <https://doi.org/10.5194/hess-17-1783-2013>.
- Parente, J., Girona-García, A., Lopes, A.R., Keizer, J.J., Vieira, D.C.S., 2022. Prediction, validation, and uncertainties of a nation-wide post-fire soil erosion risk assessment in Portugal. *Sci. Rep.* 12, 2945. <https://doi.org/10.1038/s41598-022-07066-x>.
- Pianosi, F., Sarrazin, F., Wagener, T., 2015. A Matlab toolbox for Global Sensitivity Analysis. *Environ. Model. Softw.* 70, 80–85. <https://doi.org/10.1016/j.envsoft.2015.04.009>.
- Poesen, J., Nachtergaele, J., Verstraeten, G., Valentin, C., 2003. Gully erosion and environmental change: importance and research needs. *CATENA, Gully Eros. Glob. Change* 50, 91–133. [https://doi.org/10.1016/S0341-8162\(02\)00143-1](https://doi.org/10.1016/S0341-8162(02)00143-1).
- [dataset]Portmann, F.T., 2008. Global dataset of monthly growing areas of 26 irrigated crops: version 1.0. Univ.-Bibliothek Frankfurt am Main.
- Portmann, F.T., Siebert, S., Döll, P., 2010. MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Glob. Biogeochem. Cycles* 24. <https://doi.org/10.1029/2008GB003435>.
- Qi, W., Chen, J., Li, L., Xu, C.-Y., Xiang, Y., Zhang, S., Wang, H.-M., 2021. Impact of the number of donor catchments and the efficiency threshold on regionalization performance of hydrological models. *J. Hydrol.* 601, 126680 <https://doi.org/10.1016/j.jhydrol.2021.126680>.
- Rane, N.L., Jayaraj, G.K., 2022. Enhancing SWAT model predictivity using multi-objective calibration: effects of integrating remotely sensed evapotranspiration and leaf area index. *Int. J. Environ. Sci. Technol.* <https://doi.org/10.1007/s13762-022-04293-7>.
- [dataset] RGI Consortium, 2017. Randolph Glacier Inventory - A Dataset of Global Glacier Outlines, Version 6. Natl. Snow Ice Data Cent. (<https://nsidc.org/data/nsidc-0770/versions/6>).
- Shah, S., Duan, Z., Song, X., Li, R., Mao, H., Liu, J., Ma, T., Wang, M., 2021. Evaluating the added value of multi-variable calibration of SWAT with remotely sensed evapotranspiration data for improving hydrological modeling. *J. Hydrol.* 603, 127046 <https://doi.org/10.1016/j.jhydrol.2021.127046>.
- Shin, S., Her, Y., Muñoz-Carpena, R., Khare, Y.P., 2023. Multi-parameter approaches for improved ensemble prediction accuracy in hydrology and water quality modeling. *J. Hydrol.* 622, 129458 <https://doi.org/10.1016/j.jhydrol.2023.129458>.

- Shrestha, D.P., Jetten, V.G., 2018. Modelling erosion on a daily basis, an adaptation of the MMF approach. *Int. J. Appl. Earth Obs. Geoinf.* 64, 117–131. <https://doi.org/10.1016/j.jag.2017.09.003>.
- Sidle, R.C., Jarihani, B., Kaka, S.I., Koci, J., Al-Shaibani, A., 2019. Hydrogeomorphic processes affecting dryland gully erosion: Implications for modelling. *Prog. Phys. Geogr. Earth Environ.* 43, 46–64. <https://doi.org/10.1177/0309133318819403>.
- Siebert, S., Burke, J., Faures, J.M., Frenken, K., Hoogeveen, J., Döll, P., Portmann, F.T., 2010. Groundwater use for irrigation – a global inventory. *Hydrol. Earth Syst. Sci.* 14, 1863–1880. <https://doi.org/10.5194/hess-14-1863-2010>.
- Sirisena, T.A.J.G., Maskey, S., Ranasinghe, R., 2020. Hydrological Model Calibration with Streamflow and Remote Sensing Based Evapotranspiration Data in a Data Poor Basin. *Remote Sens.* 12, 3768. <https://doi.org/10.3390/rs12223768>.
- SMHI, 2023. HYPE Model Documentation. (<http://www.smhi.net/hype/wiki/>).
- Strömqvist, J., Arheimer, B., Dahné, J., Donnelly, C., Lindström, G., 2012. Water and nutrient predictions in ungauged basins: set-up and evaluation of a model at the national scale. *Hydrol. Sci. J.* 57, 229–247. <https://doi.org/10.1080/02626667.2011.637497>.
- Theuring, P., Rode, M., Behrens, S., Kirchner, G., Jha, A., 2013. Identification of fluvial sediment sources in the Kharaa River catchment, Northern Mongolia. *Hydrol. Process.* 27, 845–856. <https://doi.org/10.1002/hyp.9684>.
- Troin, M., Martel, J.-L., Arsenault, R., Brissette, F., 2022. Large-sample study of uncertainty of hydrological model components over North America. *J. Hydrol.* 609, 127766 <https://doi.org/10.1016/j.jhydrol.2022.127766>.
- [dataset] United Nations Environment Programme, 2018. GEMStat database of the Global Environment Monitoring System for Freshwater (GEMS/Water) Programme. International Centre for Water Resources and Global Change, Koblenz. Accessed 06 November 2019. Available upon request from GEMS/Water Data Centre: gemstat.org.
- van Esse, W.R., Perrin, C., Booij, M.J., Augustijn, D.C.M., Fenicia, F., Kavetski, D., Lobligois, F., 2013. The influence of conceptual model structure on model performance: a comparative study for 237 French catchments. *Hydrol. Earth Syst. Sci.* 17, 4227–4239. <https://doi.org/10.5194/hess-17-4227-2013>.
- van Kempen, G., van der Wiel, K., Melsen, L.A., 2021. The impact of hydrological model structure on the simulation of extreme runoff events. *Nat. Hazards Earth Syst. Sci.* 21, 961–976. <https://doi.org/10.5194/nhess-21-961-2021>.
- Walling, D.E., 2005. Tracing suspended sediment sources in catchments and river systems. *Sci. Total Environ.* 344, 159–184. <https://doi.org/10.1016/j.scitotenv.2005.02.011>.
- Walling, D.E., Collins, A.L., Stroud, R.W., 2008. Tracing suspended sediment and particulate phosphorus sources in catchments (Characterization and apportionment of nutrient and sediment sources in catchments). *J. Hydrol.* 350, 274–289. <https://doi.org/10.1016/j.jhydrol.2007.10.047>.
- Wilby, R.L., 2005. Uncertainty in water resource model parameters used for climate change impact assessment. *Hydrol. Process.* 19, 3201–3219. <https://doi.org/10.1002/hyp.5819>.
- Wisser, D., Frolking, S., Douglas, E.M., Fekete, B.M., Vörösmarty, C.J., Schumann, A.H., 2008. Global irrigation water demand: Variability and uncertainties arising from agricultural and climate data sets. *Geophys. Res. Lett.* 35 <https://doi.org/10.1029/2008GL035296>.
- Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P.D., Allen, G.H., Pavelsky, T.M., 2019. MERIT Hydro: A High-resolution Global Hydrography Map Based On Latest Topography Dataset. *Water Resour. Res.* 55, 5053–5073. <https://doi.org/10.1029/2019WR024873>.
- Yang, H., Xiong, L., Ma, Q., Xia, J., Chen, J., Xu, C.-Y., 2019. Utilizing satellite surface soil moisture data in calibrating a distributed hydrological model applied in humid regions through a multi-objective Bayesian hierarchical framework. *Remote Sens.* 11, 1335. <https://doi.org/10.3390/rs11111335>.
- Yatheendradas, S., Wagener, T., Gupta, H., Unkrich, C., Goodrich, D., Schaffner, M., Stewart, A., 2008. Understanding uncertainty in distributed flash flood forecasting for semiarid regions. *Water Resour. Res.* 44. <https://doi.org/10.1029/2007WR005940>.
- Yonce, H.N., Keeley, A., Canfield, T.J., Butcher, J.B., Paul, M.J., 2019. Integration of SWAT and HSPF for simulation of sediment sources in legacy sediment-impacted agricultural watersheds. *J. Am. Water Resour. Assoc.* 55, 497–510. <https://doi.org/10.1111/1752-1688.12731>.
- Zhang, Y., Shao, Q., Zhang, S., Zhai, X., She, D., 2016. Multi-metric calibration of hydrological model to capture overall flow regimes. *J. Hydrol.* 539, 525–538. <https://doi.org/10.1016/j.jhydrol.2016.05.053>.