

Only a matter of time? Evaluating modeling strategies while building a national operational flood forecast system for Denmark

Conrad Brendel^{a,*}, Raphaël Payet-Burin^b, Phillip D. Aarestrup^{b,c},
Sanita Dhaubanjari^b, Maggie H. Madsen^b, Grith Martinsen^b,
Lucas Dalgaard Jensen^b, Jonas W. Pedersen^b, Charlotte A. Plum^b,
Emma D. Thomassen^b, Cecilie Thrysøe^b, René Capell^a, Jafet C.M. Andersson^a,
Michael B. Butts^{b,d}

^a Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

^b Danish Meteorological Institute, Flooding and Hydrology, Copenhagen Ø, 2100, Denmark

^c Technical University of Denmark, DTU Sustain, Kgs. Lyngby, 2800, Denmark

^d Lund University, Faculty of Engineering, Division of Water Resources Engineering, Lund, 221 00, Sweden

ARTICLE INFO

Keywords:

Flood early warnings
Operational forecasting
Hydrological modeling strategy
Development time
HYPE
LSTM
Denmark

ABSTRACT

In January 2022, the Danish Meteorological Institute was appointed the national flood warning authority for Denmark and tasked with developing an operational flood warning system within three years. Three different model development strategies were pursued in parallel to set up hydrological models for river discharge predictions, providing a unique real-world case study to compare their merits. A “top-down” model for Denmark was extracted from the process-based, pan-European E-HYPE model developed from large-scale data sources. A second process-based model, DK-HYPE, and a data-driven Long Short-Term Memory (LSTM) model were developed from the “bottom-up” using national data sources and high-resolution catchment delineations. This study evaluates the strategies not only in terms of classical model performance, but also by invested development time and operational benefits. Results show that the top-down approach enabled early operational readiness (1 day of work) and rapid development of operational workflows but was limited in its overall performance (Median KGE = 0.50). The bottom-up process-based approach required significantly longer development efforts (19–32 months) but provided better model performance through local data and improved process representation (KGE = 0.81). The data-driven approach gave the best model performance (KGE = 0.91) with a relatively short development time (3–5 months). Despite lower overall performance, the top-down model proved competitive during two investigated flood events, highlighting its value in early system development. These findings demonstrate the benefits of a multi-model, hybrid development strategy and provide practical guidance for the establishment of flood warning systems, including novel estimates of the time required for model development and operationalization.

* Corresponding author.

E-mail addresses: conrad.brendel@smhi.se (C. Brendel), rpb@dmu.dk (R. Payet-Burin), phaa@dmu.dk (P.D. Aarestrup), sdh@dmu.dk (S. Dhaubanjari), mma@dmu.dk (M.H. Madsen), gma@dmu.dk (G. Martinsen), ldj@dmu.dk (L.D. Jensen), jope@dmu.dk (J.W. Pedersen), cpl@dmu.dk (C.A. Plum), emt@dmu.dk (E.D. Thomassen), cet@dmu.dk (C. Thrysøe), rene.capell@smhi.se (R. Capell), jafet.andersson@smhi.se (J.C.M. Andersson), mbu@dmu.dk (M.B. Butts).

<https://doi.org/10.1016/j.ijdr.2026.106393>

Received 1 May 2026; Received in revised form 14 August 2026; Accepted 20 August 2026

Available online 20 August 2026

2212-4209/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Abbreviations	
CSI	Critical Success Index
DMI	Danish Meteorological Institute
KGE	Kling-Gupta Efficiency
NSE	Nash-Sutcliffe Efficiency
LSTM	Long Short-Term Memory
SLC	Soil type - Landuse Combination
SMHI	Swedish Meteorological & Hydrological Institute

1. Introduction

Floods are the most frequent type of natural disaster and also among the costliest and devastating. From 2000 to 2019, 44% of global disasters were associated with floods, and an average of 163 flood-related disasters are reported each year [1]. Floods have considerable financial impacts, with average annual losses associated with flooding totaling USD 388 billion globally [2]. The human impact of flooding is also significant, and floods accounted for 9% of disaster-related deaths from 2000 to 2019 [1]. Flood risks and impacts are intensifying, reflecting both increasing hazard frequency and growing societal vulnerability. From 2000 to 2020, the number of recorded flood-related disasters increased by 134% compared with the previous two decades, and from 1970 to 2020 the number of people exposed to floods increased by 24.9% [2]. Consequently, the financial and human costs associated with flooding are projected to rise considerably in the near future.

As weather and climate conditions become more extreme, the need for effective flood warning systems is greater than ever. In 2022, the UN launched the Early Warnings for All initiative, which aims to ensure that everyone on earth is protected by early warning systems by 2027. Despite a 113% increase in the number of countries reporting a capability for multi-hazard early warning systems from 2015 to 2024, nearly 40% of countries (76 total) still do not report having such a warning system [3]. Flood hazard early warning systems currently in operation span a range of scales from global systems such as GloFAS [4] and Google Flood Hub [5], continental systems such as EFAS [6], regional systems such as FANFAR [7], a multitude of state and national systems [8–13], to even local systems such as MERLIN [14], MIDAS [15], SLAPIS [16], and PUFFIN [17].

Hydrological predictions have traditionally been performed using process-based models such as HYPE, but recent studies have found the performance of data-driven models to match or even surpass that of established process-based models [18–20]. Whereas process-based models use pre-defined structures, variables, and parameters to emulate our understanding of hydrological processes, data-driven models focus on input-output relationships and can be trained without knowledge of the physical processes in the model domain [21]. In comparison to process-based models, data-driven models can also significantly accelerate the computational process by substituting iterative processes, such as the integration of differential equations, with parallel operations such as the propagation of neural networks [22]. This makes data-driven approaches well-suited for flood forecasting, and data-driven models have been applied to flood forecasting from catchment to global scales [5,20,22–25]. Criticisms of data-driven models include that data-driven models may perform more poorly than process-based models when training data is limited and that the “black-box” approach used in many data-driven models offers little insight into the uncertainty and physical interpretability of watershed processes which may be relevant when making decisions within an operational flood forecasting context [18,21,26].

While numerous studies have assessed hydrological models and their flood forecasting performance across various model concepts, there is limited practical guidance for new flood warning authorities on how to establish flood forecasting systems from scratch or on the resources required to develop and operationalize them. The process of model development and operationalization of large-scale flood forecasting systems has been described for a few large-scale flood warning systems, e.g. EFAS [6] and GloFAS [4]. Stephens and Cloke [13] discuss how institutional challenges and scientific development affect the societal benefits of the flood warnings from the UK Flood Forecasting Centre. More challenging than developing a hydrological forecasting model is efficiently using the available resources to construct the entire operational flood warning chain including the acquisition and processing of input data, model development, delivery of results, issuance of warnings, and training of the operational forecasting staff. To our knowledge, no study has explicitly quantified the time and resource requirements of different model development approaches to identify the most effective pathway for rapid development of an operational forecasting system. Therefore, clear guidance on how to efficiently develop flood forecast models is essential to support initiatives such as Early Warnings for All and to lower barriers to the implementation of flood early warning systems.

In this paper, we compare different hydrological modeling strategies and their resource requirements within the context of national scale operational flood forecasting in Denmark. Denmark is especially vulnerable to acute flooding from storm surges, summer cloudburst events and prolonged periods of rain in the winter season, and the risk of flooding in the country is projected to increase due to urban development in low-lying areas and the impact of climate change [27–29]. Although prior to 2022, the Danish Meteorological Institute (DMI) provided warnings to the public for extreme rainfall and storm surges, there was no national authority for flood early warning in Denmark. Instead, management of flood risks and performing preventative emergency service actions was the sole responsibility of local municipalities. Motivated by the need to respond effectively to the growing risks of flooding, DMI was appointed in 2022 by the Danish government as the national authority for flood early warning with a challenging deadline to develop an operational hydrological warning service — covering floods from sea level rise, cloud bursts as well as riverine floods — by 2025. The primary goal

of this warning service is to determine the risk of flooding and identify threatened areas while allowing time for emergency services and citizens to respond effectively. Prior to developing their flood forecasting models, DMI based river forecasts on EFAS products and a historical analysis of recent flooding events. In this paper, however, we focus solely on the modeling efforts supporting warnings for riverine floods. The three conditions — vulnerability to flooding, no prior flood early warning system and a strict time constraint — created a natural testing environment in Denmark for evaluating different hydrological modeling strategies in parallel under real-world constraints, including trade-offs between performance, development time, and operational readiness.

While developing their operational hydrological warning service, DMI decided on a strategy in which two modeling approaches, “top-down” and “bottom-up”, were tested during different stages of the project. A top-down approach uses an existing large-scale model to quickly obtain a first operational model and to inform the development of an adapted model by e.g. extracting the desired domain from the large-scale model and then refining the model structures, process descriptions, and calibration. The top-down model used here is the process-based HYPE hydrological model for Denmark, hereafter referred to as “E-HYPE (DK)”, that was extracted from an existing pan-European HYPE model setup [30]. Extracting a model for Denmark using this top-down approach allowed DMI to exploit the opportunity for rapid deployment of a national warning system by developing the warning system infrastructure and forecast chain early in the project using an existing model. Furthermore, the E-HYPE (DK) model provided experience to inform and direct further modeling efforts. Later, DMI built a separate process-based HYPE hydrological model, called “DK-HYPE”, from the bottom-up (i.e. constructed entirely from scratch) using national data sources and catchment delineations not available at European and global scales. This model allowed DMI to obtain more localized forecasts and higher spatial resolution forecasting capabilities than would be possible with E-HYPE (DK). Finally, DMI constructed a data-driven Long Short-Term Memory (LSTM) hydrological model, “DK-LSTM”, from the bottom-up using the same datasets as their “DK-HYPE” process-based model. The DK-LSTM model allowed DMI to evaluate state-of-the-art data-driven forecasting methods in comparison to traditional process-based models. Combined, the three different models allowed DMI to take advantage of the benefits of multi-model forecasting [31–33].

The three national-scale flood forecast models developed for the operational flood warning service for Denmark present a unique opportunity to compare top-down (E-HYPE (DK)) versus bottom-up (DK-HYPE & DK-LSTM) modeling approaches and process-based (E-HYPE (DK) & DK-HYPE) vs. data-driven (DK-LSTM) hydrological models for a large-scale modeling domain. There is a trade-off between the time required to develop a national hydrological model and the additional information gained from using multiple model forecasts. In this study, we explicitly quantify person-time invested vs. forecasting value across the various model paradigms and reflect on our experiences and lessons learned in relation to the following questions regarding operational flood forecasting at the national scale:

1. How do different hydrological model development strategies (top-down vs. bottom-up; process-based vs. data-driven) compare in terms of performance and time to operational readiness in a national flood forecasting context?
2. How do the models perform for selected historical flood events, and what levels of model performance and physical explainability are sufficient for issuing flood warnings?
3. What institutional and practical lessons can be drawn from the development process, including strengths and limitations of the evaluated modeling strategies, which may inform similar efforts by other national flood warning authorities?

2. Data & methods

2.1. Hydrological model software

2.1.1. HYPE (process-based)

Hydrological Predictions for the Environment (HYPE) is an open source, process-based, semi-distributed hydrological model developed by the Swedish Meteorological and Hydrological Institute to simulate water quantity and quality [34]. In HYPE, drainage areas are represented as subbasins characterized by hydrological response units called SLCs (Soil type – Landuse Combination classes), which represent the unique combinations of physical attributes such as landuse/land cover, soil type, elevation, and crop cover. HYPE provides routines for most major land surface and subsurface processes including snow/ice melting and accumulation, evapotranspiration, surface and macropore flow, soil moisture and groundwater fluctuation, aquifer recharge/discharge, irrigation and abstractions, runoff and discharge generation, and routing through rivers, lakes, and reservoirs. These routines are governed by both general parameters (those which have a single value across the entire model domain) and parameters tied to specific physical attributes (typically soil or landuse), which allow for spatial variations in the model behavior. Model simulations are typically performed at the daily time step with outputs aggregated across the subbasin area or summarized for the subbasin outlet. Complete documentation of the HYPE model and its routines can be found in Ref. [34].

2.1.2. CudaLSTM (data-driven)

CudaLSTM is a specialized Long Short-Term Memory (LSTM) model implemented within the open-source Python-based Neural Hydrology (<https://neuralhydrology.readthedocs.io/en/latest/>) framework. An LSTM model is a type of recurrent neural network (RNN) model which solves the “vanishing-gradient” problem occurring in traditional RNNs [35]. LSTM models work with a memory-mimicking cell-structure including both an input, output and forget gate, which allows the model to capture both long-term dependencies (e.g. groundwater) as well as short-term responses (e.g. runoff). In addition to the suitability of LSTM models for predicting hydrological variables, it is also possible to perform regional modeling with a single LSTM model by training a single LSTM model across multiple catchments and utilizing dynamic forcings for time-dependent dynamics and static attributes for generalization

across the catchments. This approach for solving hydrological problems was demonstrated in the foundational work of Kratzert et al. [36,37]. The value of using an LSTM model in predicting streamflow has since been demonstrated by numerous use-cases showing similar performance as process-based hydrological models (e.g. Ref. [37–39]).

2.2. Hydrological model setups

Three separate hydrological model setups were developed for Denmark during the project — the top-down “E-HYPE (DK)” and bottom-up “DK-HYPE” process-based HYPE models and the bottom-up data-driven “DK-LSTM” model — each with different versions as the models developed (Fig. 1). A general overview of the setup and calibration of these models is provided in Sections 2.2.1 - 2.2.3, but additional details are provided in the Supplemental Information Section S1. “Operational readiness” is a term we use throughout the paper. We deemed models to be “operationally ready” if they were sufficiently calibrated, validated, and tested to be put into operational production. The E-HYPE (DK) model was considered operationally ready from the start of the project since the full pan-European E-HYPE model setup was already in production at SMHI. Our criteria for the Danish national hydrological models (DK-HYPE and DK-LSTM) to be deemed operationally ready were to (1) spatially cover the most flood prone regions of Denmark with an improved hindcast KGE score compared to the E-HYPE (DK) models and (2) produce hydrographs with correct dynamics for the catchments upstream of flood prone regions. The model versions which were not classified as operationally ready were never run in forecast mode.

2.2.1. E-HYPE (DK)

At the start of the project, the “top-down” E-HYPE (DK) model was created by extracting all Danish subbasins from the pan-European E-HYPE hydrological model developed by SMHI [30] and then performing a recalibration focused explicitly on model performance for Denmark. The E-HYPE model (version 4) was developed using a variety of large-scale open-source datasets (Table 1), and the full pan-European model setup contains 42,629 subbasins (median area = 170 km²) based on MERIT Hydro [40] DEM and flow direction/accumulation grids with a river network, lake polygons, and subbasin outlet points derived from the EU-Hydro dataset [41]. In the E-HYPE (DK) model extraction, there are 241 subbasins with areas ranging from 0.18 to 864 km² (median = 173 km²). Meteorological forcing data is from HydroGFD3.2 [42], and the stream flow calibration data is from the Global Runoff Data Centre [43]. The initial extraction of the E-HYPE (DK) model (from E-HYPE v4.1.1) took approximately one day, which allowed DMI to start working on the development of the operational forecasting chain immediately at the beginning of the project. Later in the project, the E-HYPE (DK) model setup was updated to incorporate developments from the full E-HYPE model (version 4.3.0) which included changes to the snowmelt routine and updated calibration parameter ranges and sampling distributions.

Both pan-European E-HYPE model setups (v4.1.1 and v4.3.0) were calibrated by running the models on a supercomputer (using 16 cores for approximately 24 h) using ensembles of parameter sets generated with the Latin Hypercube sampling method within the SAFE toolbox [58]. This method maintains a record of the model performance at each calibration station for each parameter set, and

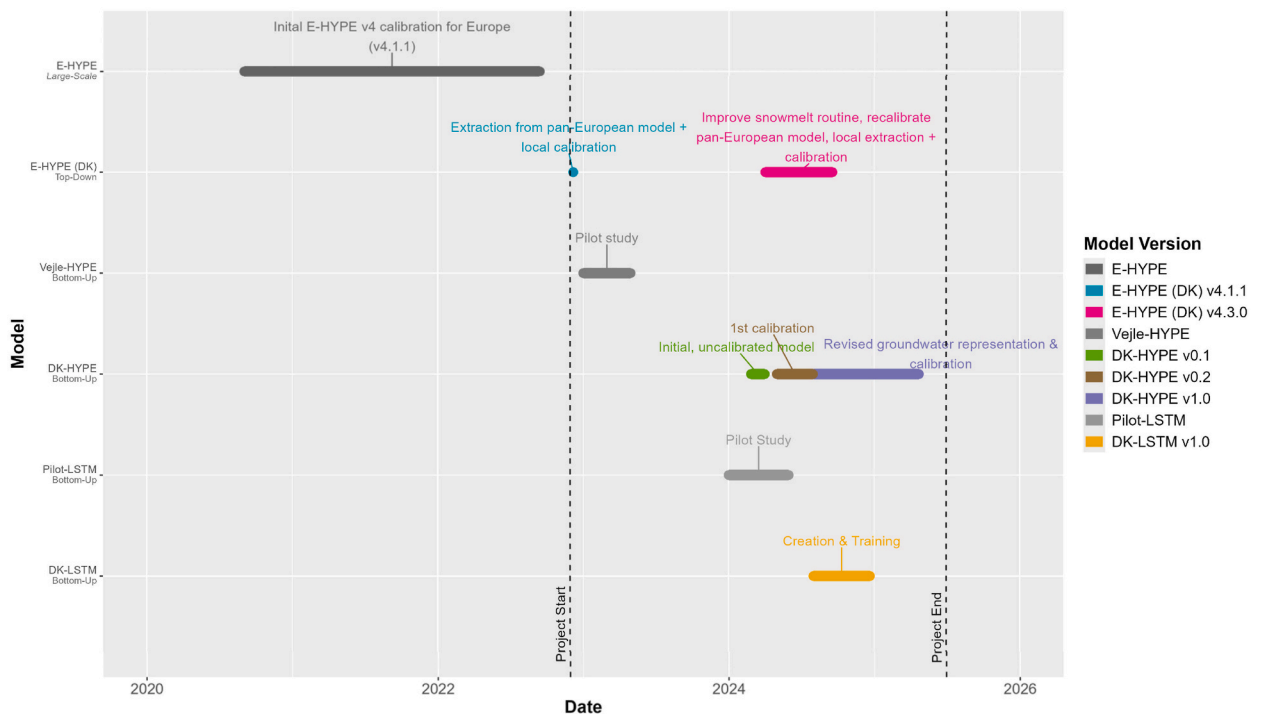


Fig. 1. Model development timeline.

Table 1
Datasets used in the development of the hydrological models.

Category	Data	E-HYPE (DK)		Reference
		Source	Link	
Model Domain & Routing	Model Domain & Subbasin Delineation	MERIT Hydro	http://hydro.iis.u-tokyo.ac.jp/~yamada/MERIT_Hydro/index.html	[40]
	River Network, Lake Polygons, & Outlet Points	EU-Hydro	https://land.copernicus.eu/imagery-in-situ/eu-hydro/eu-hydro-river-network-database	[41]
Subbasin Description	Landuse	CORINE	https://land.copernicus.eu/pan-european/corine-land-cover/clc-2006	[44]
	Glaciers	ESA CCI	http://maps.elie.ucl.ac.be/CCI/viewer/download.php	[45]
	Climate Zones	RGI	https://www.glims.org/RGI/	[46]
	Soils	Köppen-Geiger	http://koeppen-geiger.vu-wien.ac.at/present.htm	[47]
		ESDB	https://esdac.jrc.ec.europa.eu/	[48]
		DMSW	https://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/faunesco-soil-map-of-the-world/en/	[49]
	Crops	NUTS	https://ec.europa.eu/eurostat/web/nuts/background	[50]
Water Management	Management & Irrigation	Various National Sources		
		GMIA	http://www.fao.org/nr/water/aquastat/irrigationmap/index.stm	[51]
		MIRCA	https://www.uni-frankfurt.de/45218023/MIRCA	[52,53]
	Point Sources	UWWTD	https://www.eea.europa.eu/en/datahub/datahubitem-view/7c22ae09-4827-4a7b-a190-1504a22cf094#tab-additional-information	
		HydroWaste v1.0	https://www.hydrosheds.org/products/hydrowaste	[54]
Category	Data	DK-HYPE & DK-LSTM		Reference
		Source	Link	
Calibration & Validation Data	Discharge	GRDC	https://portal.grdc.bafg.de/applications/public.html?publicuser=PublicUser#dataDownload/Home	
Model Domain & Routing	Model Domain, Subbasin Delineation, River Network	DCE AU, Vandskel, ID15oplande	https://arealdata.miljoeportal.dk/datasets/urn:dmp:ds:oplande-vandskel	[55]
	Lake Polygons	DCE AU, vp3, 2021	https://arealdata.miljoeportal.dk/datasets/urn:dmp:ds:karakterisering-typologi-soeer-vandomraadeplaner-2021-2027	[55]
Subbasin Description	Landuse	DCE AU, Basemap04, 2021	https://envs.au.dk/om-instituttet-1/faglige-omraader/samfund-miljoe-og-ressourcer/arealanvendelse-og-gis/basemap	[56]
	Soils	DCE AU, 2013	https://dca.au.dk/forskning/den-danske-jordklassificering/https://acsess.onlinelibrary.wiley.com/doi/10.2136/sssaj2012.0275	[57]
Calibration & Validation Data	Discharge, Outlet Points	SGAV discharge databases	Samling af Vandløbsdata (SAV) Vanddatabasen https://vandah.miljoeportal.dk/api/swagger/index.html	

this enables different selection criteria (e.g. stations and statistics) to be used without requiring the models to be rerun. Parameter sets for the E-HYPE (DK) v4.1.1 and v4.3.0 model setups were selected from their respective parameter ensembles to optimize the Kling-Gupta Efficiency (KGE) [59] model performance for discharge in Denmark. Average model run times for the full E-HYPE pan-European domain are 8–10 h for a 30-year simulation whereas run times for the E-HYPE (DK) domain for the same simulation period take approximately 2 min. Additional details regarding the E-HYPE model setup and calibration can be found in Ref. [30] and in the Supplemental Information Section S1.

2.2.2. DK-HYPE

Development of the “bottom-up” DK-HYPE model commenced during the second year of the project. This model has 3447 catchments ranging from 0.02 to 106 km² (median = 12.2 km²) based on the Danish national catchment delineation. The precipitation, temperature, and potential evapotranspiration (PET) forcing are from DMI's climate dataset, ClimateGrid, in 1 × 1 km resolution (PET is calculated using the modified Makkink method). Description of the dataset and open coarser resolution data are accessible through DMI Open Data [60]. The stream flow data used for calibration is from the Danish Agency for Green Transition and Aquatic Environment. Many Danish streams are heavily influenced by groundwater, so unlike the E-HYPE (DK) models, DK-HYPE makes use of the HYPE aquifer module [34]. The aquifer module is simply a linear reservoir, which sits beneath the soil layers. It receives water (aquifer recharge) from the bottom soil layer and returns water directly to the river (aquifer return flow). Both recharge and return flow are determined by linear reservoir time constants, but the return flow can be multiplicatively scaled with a parameter. In DK-HYPE, all subbasins have their own aquifer with which they vertically exchange water.

DK-HYPE was set up in an iterative process, in which three model versions were completed during the project. Prior to the development of DK-HYPE, a pilot study model (Vejle-HYPE, Fig. 1) was set up and calibrated for a smaller area around the city of Vejle in Denmark [61]. The tools created and the experience gathered in setting up Vejle-HYPE laid the foundation for the development of the different DK-HYPE versions:

- DK-HYPE v0.1: The very first version of DK-HYPE is 0.1. This model setup is uncalibrated; the parameters were set as a mix of inherited values from E-HYPE and the Vejle-HYPE models.
- DK-HYPE v0.2: The first calibrated version of DK-HYPE.
- DK-HYPE v1.0: In this version, the groundwater representation was revised by calibrating the aquifer return flow parameters (delay constant and the multiplicative scaling factors) to improve stream baseflow. These parameters were fitted to each calibration station individually, with all subbasins upstream of a calibration station receiving the same calibrated parameter value. Lakes were also included in the model for the first time and were implemented as outlet lakes using a universal lake depth and rating curves for all lakes above 1 km².

Calibration of the DK-HYPE v0.2 and v1.0 models was performed by wrapping the HYPE model setups inside the OSTRICH optimization toolbox, which is a flexible, model-agnostic tool [62]. OSTRICH's Dynamically Dimensioned Search (DDS) algorithm was used during the calibrations to optimize the average KGE score across 37 calibration stations spread across the country. DK-HYPE v0.2 and v1.0 shared identical calibration settings, except for the aquifer parameters. Individual DK-HYPE model runs took approximately 10–15 min, while OSTRICH DDS calibration runs required around 1–1.5 days for 150 iterations on a standard computer. The computation time for DK-HYPE was primarily influenced by the number of HYPE output files requested in the model run while that for OSTRICH was tied to the number of iterations allowed. OSTRICH calibrations, especially some tests with higher iterations, were typically executed on DMI's servers, although they did not need advanced computational resources. Additional details regarding the DK-HYPE setup and calibration can be found in the Supplemental Information Section S1.

2.2.3. DK-LSTM

A purely data-driven LSTM model, DK-LSTM, was set up for Denmark towards the end of the project (Fig. 1) using the open-source Python based CudaLSTM model framework from Neural Hydrology (<https://neuralhydrology.readthedocs.io/en/latest/>). The DK-LSTM model uses the same resolution and catchment delineation as the process-based hydrological model DK-HYPE and has data for 570 gauged catchments. The dynamic data consists of catchment averaged daily time series of meteorological forcings — including precipitation, potential evapotranspiration, and temperature — based on the 1 × 1 km gridded ClimateGrid dataset from DMI [60]. The static attributes were compiled specifically for the DK-LSTM model and consisted of catchment statistics from gridded datasets describing landuse [56] (catchment share of urban areas, agriculture, forest, “dry” nature, “wet” nature”, lakes and wetlands, respectively), soil types [57] (percentage of clay, fine sand and coarse sand across four soil profiles at depths of 0–30, 30–60, 60–90 and 90–120 cm), topography based on the Danish national Digital Terrain Model (DTM) in a 1 m resolution [63] (minimum, maximum, mean and median elevation and slope, as well as percentage flat area with slopes smaller than 3 m/km) and climatology calculated from DMI [60] (long-term yearly mean temperature, potential evapotranspiration and precipitation). DK-LSTM was not based on existing national CAMELS datasets for Denmark because we wanted a model set-up which we could maintain ourselves and we wanted

Table 2

Time estimates for model development.

	E-HYPE (DK) v4.1.1	E-HYPE (DK) v4.3.0	DK-HYPE v0.1	DK-HYPE v0.2	DK-HYPE v1.0	DK-LSTM v1.0
Description	Extraction of Danish submodel from pan-European E-HYPE model and identification of best parameter set for Denmark	Improvements to snowmelt routine, recalibration of pan-European E-HYPE model setup, extraction of Danish submodel, and identification of best parameter set for Denmark	Initial, uncalibrated DK-HYPE model setup	First calibration of DK-HYPE model	Revised groundwater representation and calibration of groundwater aquifer parameters	Creation and training of model
Operationally Ready?^a	Yes	Yes	No	No	Yes	Yes
Period Start	2022-12-07	2024-04-02	2024-02-26	2024-05-01	2024-08-01	2024-08-01
Period End	2022-12-07	2024-09-18	2024-04-01	2024-07-31	2025-04-23	2024-12-21
Period Length [Period End – Period Start]	1 day	5.5 Months	1.25 Months	3 Months	8.75 Months	4.75 Months
Number of Team Members	1	2	6	7	9	3
Development Time (Person-Months)	0-1	2-6	2-4	6-10	11-18	3-5
Cumulative Development Time (Person-Months) [Development Time + Cumulative Development Time for Previous Versions]	0-1	2-7	2-4	8-14	19-32	3-5

^a Operationally ready models were those considered sufficiently calibrated, validated, and tested to be put into operational production (See Section 2.2).

a higher data coverage of time-series data than what was found in Koch and Schneider [39].

The DK-LSTM model was trained against time series of observed daily discharge for the 570 gauged catchments. In data-driven models, observations are split into training, validation, and test sets so that the model can be tuned while keeping a completely untouched dataset (the test set) for an unbiased final evaluation. Training of the DK-LSTM model was performed by splitting each individual time series into sets with 70% of the data for training, 15% for validation, and 15% for testing. The data selected for training was the only data seen by the model during training of the neural network, while the validation data was used for evaluating aspects like convergence and overfitting at every epoch. The test data served as an independent dataset for evaluating model performance. The split was based on the first and last data point in each time series, and it was ensured that there was no data leakage between splits. NSE was chosen as the loss function which the model was trained against. The model was allowed to train for 30 epochs with a stepwise decreasing learning rate. The sequence length was selected to 365 days, and training was performed with a batch size of 256. The LSTM model had a hidden size of 64 and a 0.4 dropout. Hardware used for training the DK-LSTM model was NVIDIA GeForce RTX 4060 Laptop GPU with a training time of 1.5 min for each of the 30 epochs.

2.3. Time estimates for model development & composition of development teams

Model development was carried out by teams of hydrologists at SMHI (E-HYPE/E-HYPE(DK) models) and DMI (DK-HYPE and DK-LSTM models). Estimates of the time invested in each model setup are presented in Table 2, and the composition of the development teams are summarized in Sections 2.3.1–2.3.3. The time that each employee spent during the model development phase could not be calculated exactly because team members had other parallel work tasks during the project period and could not dedicate 100% of their working time to model development. Furthermore, the number of available team members varied between different model versions. Thus, the time invested in each model was estimated retrospectively but represents our best-effort expert judgement.

Time estimates for model development were based on (1) the period of model development of each model version, (2) the number of people contributing, and (3) an estimate of the relative amount of time each person contributed with. Git was used for version control during model development, and Git tags along with Git commits were used to determine the model development period for each version. For each model version, the number of contributors involved in the development was quantified, and all contributing employees were asked to inspect their personal calendars and quantify their work availability in units of “full-time person-months”. Here, work availability was defined as the total length of the development period minus their personal periods of absence (holidays/vacation, sick leave, parental leave, part-time employment, etc.). The personal work availabilities were then summed to a combined team availability, which served as an upper limit for the development time if all relevant employees had worked only on model development. The team then evaluated the other tasks besides model development that were performed during the period and defined both their best estimate of model development time as well as reasonable minimum and maximum ranges (these are provided in Table 2 and a worked example for DK-HYPE v1.0 is provided in Section S2 in the Supplemental Information). Finally, the cumulative development time for each model version was calculated by adding the time spent for each version to the time spent on any previous model versions. Model development times could vary significantly across other organizations with different baseline capabilities (e.g. access to existing models, computational resources, staff expertise, amount of annual leave, etc.).

2.3.1. E-HYPE (DK)

Setup of the base pan-European E-HYPE model [30] was performed in 2020 by two team members from SMHI, one with a M. Sc. and one with a Ph.D. in hydrology. Calibration and development of E-HYPE v4.1.1 was performed part-time from 2020 to 2022 (prior to the start of this project) by a post-doc at SMHI with a Ph.D. in civil engineering; this post-doc had prior hydrological modeling experience, but no prior experience with HYPE specifically. Development of E-HYPE v4.3.0 took place part-time over 2023–2024 as part of this project by the same person with assistance from a coworker with a M. Sc. in engineering. Extraction of the E-HYPE (DK) models from the pan-European E-HYPE model setups and identification of the best parameter sets for Denmark was also performed by the same SMHI employee.

2.3.2. DK-HYPE

DK-HYPE was developed by the hydrology unit at DMI, which grew from 6 to 9 hydrologists during the period of the model development (Feb 2024 to April 2025). Initial experience with the HYPE software in the unit was primarily limited to the Vejle-HYPE pilot model setup implemented before the actual development of DK-HYPE (Fig. 1). During the pilot study, the DMI team developed its own in-house Python toolbox for setting up HYPE models, including data processing pipelines for HYPE's hydrological response unit definition, parsing and formatting forcing and river discharge observations, and basic tools for assessing simulation outputs and performance. Developing these generic tools required substantial effort and accelerated the later development of DK-HYPE models. Initial development of the Python toolbox is not counted in the model development time in this study, since it was completed in a separate project. However, improvements to it during DK-HYPE developments are included as they were required to set up the models. Many of the functionalities of the toolbox are available for others through SMHI's publicly available HYPE setup and analysis tools — e.g. the WHIST GIS toolbox [64] and the HYPETOOLS R package [65] — so others do not necessarily need to code their own tools. All hydrologists had a M.S. in civil engineering, and 7 out of 9 had a Ph.D. within relevant fields such as hydrological modeling and extreme rainfall. Years of experience ranged from 0 to 35+ years, while most hydrologists at DMI had 7–9 years of experience (Ph.D. included). None of the hydrologists worked full time on model development, and the primary hydrologists working on model development varied between model versions.

2.3.3. DK-LSTM

DK-LSTM was primarily developed during a university student project with DMI as the main supervisor and data provider. The team contributing to the development of DK-LSTM consisted of a bachelor level student with a bachelor's degree specialization in AI/ML from a technical university. The two team members from DMI were a Ph.D. student in ML and a Ph.D. in hydrological modeling with approximately 8 years of experience including the Ph.D. study. Overall, the student project was a 10 ECTS credit project, and DMI staff invested time in developing the full dataset for the DK-LSTM model as well as supervision of the student carrying out the project of developing and training the DK-LSTM model. In addition to the time invested by DMI staff involved in the DK-LSTM development, the

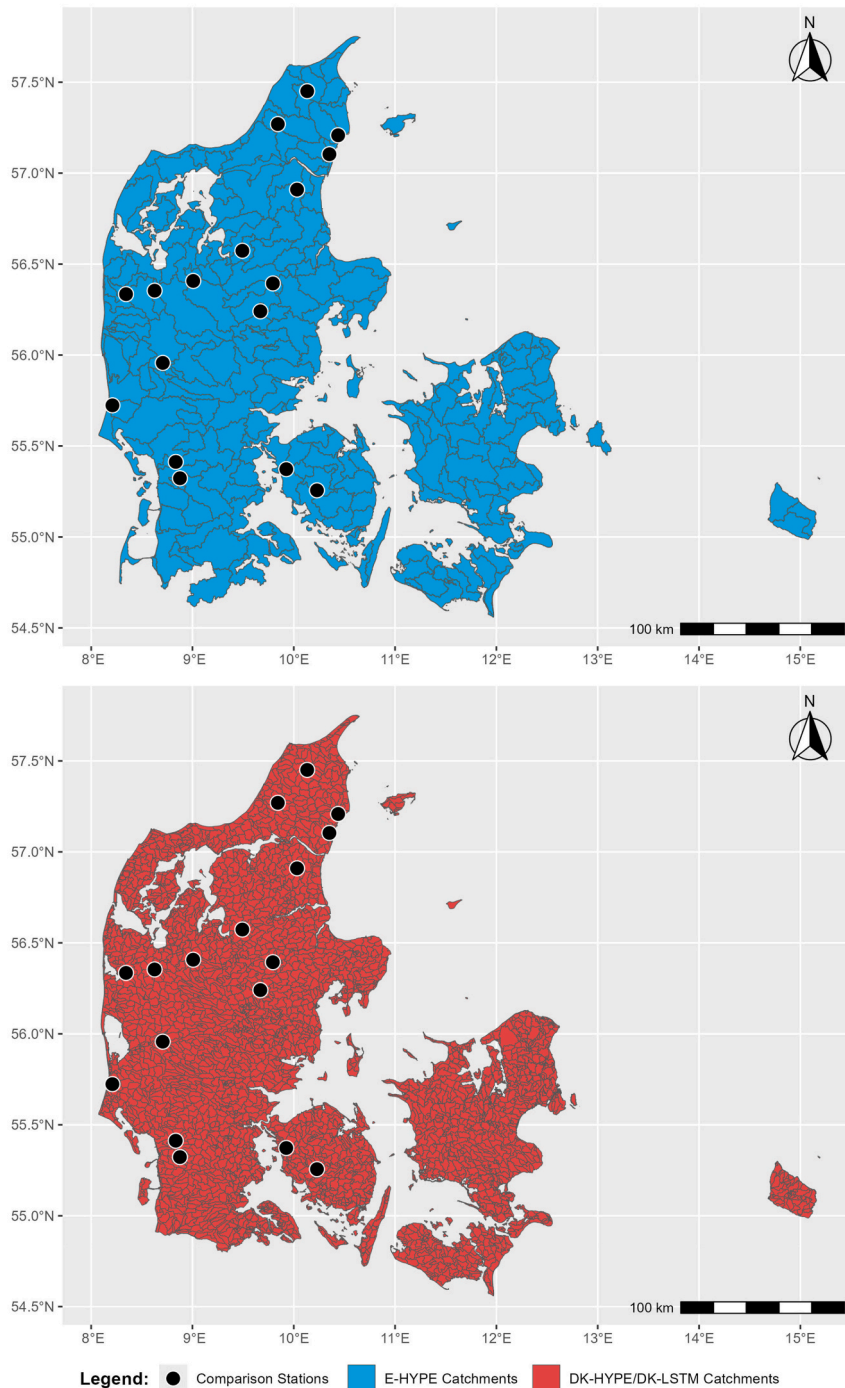


Fig. 2. Discharge stations (black circles) used for model performance comparison with catchment delineations for E-HYPE (top) and DK-HYPE/DK-LSTM (bottom).

time of the student was estimated based on an ECTS equivalent of 28 h/ECTS. Although the DK-LSTM model used several of the same national data sources as the DK-HYPE models, e.g. forcing data and local geographic properties, separate data processing pipelines were required for formatting the data to the Neural Hydrology framework. There was therefore limited reuse of HYPE-specific data processing workflows. As for DK-HYPE, generic software tools were developed in the “Pilot-LSTM” project (Fig. 1). The development of these were not counted here, as only subsequent improvements during national-scale DK-LSTM model development were included.

2.4. Model comparison

Model performance and hydrological response behavior were compared for each version of the three model setups:

- E-HYPE (DK) version v4.1.1 (the initial extraction and calibration for Denmark) and v4.3.0 (initial pan-European scale recalibration and new snowmelt routine followed by local calibration for Denmark)
- DK-HYPE versions v0.1 (initial uncalibrated model), v0.2 (first calibrated model), and v1.0 (recalibrated model with revised groundwater representation)
- DK-LSTM version v1.0 (final operational version)

Temperature and precipitation forcing data for the E-HYPE (DK) models were obtained from the same global hydrological forcing dataset, HydroGFD3.2 [42], that was used for calibration of the model. In contrast, the DK-HYPE and DK-LSTM models were run using the Danish-specific ClimateGrid meteorological forcing dataset. All of the models were run for the same simulation period and evaluated for the period 2014–2019 using flow observation data for 17 stations. There are about 570 gauged stations managed by the Danish Environmental Agency with data between 2001 and 2025, but these 17 stations were selected to provide an overlap between the catchment outlets of all models and sufficient data coverage of discharge time series for evaluation (Fig. 2). The model comparison period included large flood events in December 2015 and February 2020 and covers a wide range of hydrological conditions across Denmark, including catchments with contrasting geology, topography, groundwater influence, and runoff response (Soil and landuse maps for Denmark are included in Figure S1 and Figure S2 in the Supplemental Information). Agreement between the simulated and observed discharge timeseries was assessed using the KGE model-fit statistic.

As this study is a retrospective assessment of successive operational model developments rather than a controlled benchmarking experiment, the evaluation is not fully independent of model calibration. Specifically, DK-HYPE v0.2 and v1.0 were calibrated using 9 of the 17 evaluation stations for the whole evaluation period; the LSTM model was trained using data from all stations with a (random) 30% temporal hold-out; the E-HYPE (DK) v4.1.1 was not calibrated on any of the stations but the best parameter set for Denmark from the (calibrated) parameter ensemble was selected using all 17 stations; the E-HYPE (DK) v4.3.0 used 7 of the stations for calibration and 10 of the stations for selecting the best parameter set among the calibrated ensemble (Table 3).

To complement the KGE assessment and provide an evaluation that was independent of the calibration period, we also evaluated the models using a peak-flow performance metric over an independent validation period (2023–2024) which was not used for calibration of any model. Peak-flow performance was assessed based on the models' ability to predict streamflow events exceeding a 2-year return period threshold. For these events, we computed the hit rate (the proportion of observed exceedances that were correctly predicted by the model), the success ratio (the proportion of predicted exceedances that were actually observed), and the Critical Success Index (CSI), which combines these two measures into a single indicator of event prediction skill. The 2-year return period threshold was chosen because it represents a high flow, but is frequent enough to make reliable statistics. Return period thresholds were estimated independently for the observations and each model simulation by station. This approach evaluates the models' ability to predict the occurrence of high-flow events, independent of systematic biases in simulated discharge magnitude.

Table 3
Calibration and comparison stations for each model.

Model	Total calibration stations	Calibration stations within Denmark	Comparison stations used in calibration (out of 17)	Comparison stations used in parameter set selection for Danish domain (out of 17) ^a
DK-HYPE v0.1	0	0	0	NA
DK-HYPE v0.2	37	37	9	NA
DK-HYPE v1.0	37	37	9	NA
DK-LSTM v1.0	570	570	17	NA
E-HYPE v4.1.1	226	0	0	17
E-HYPE v4.3.0	745	19	7	10

^a Identification of the single parameter set providing the optimal performance across the Danish domain was performed via a rescaled ranking of parameter ensemble members according to both calibration and validation performance (KGE & relative error) using the method described in Ref. [30].

3. Results

3.1. Comparison of model performance

The performance between the different model versions was compared using maps (Fig. 3) and cumulative distribution functions (Fig. 4). Each version of the three model setups was included in the analysis: (1) E-HYPE (DK) v4.1.1 and v4.3.0, (2) DK-HYPE v0.1, v0.2, and v1.0, and (3) DK-LSTM v1.0. In Fig. 3, the colors represent the KGE between the simulated and observed discharges across the 17 comparison stations for the 2014–2019 evaluation period. For each point in Fig. 4, the x-axis value represents the KGE for a particular station, and the y-axis value indicates the proportion of stations with a KGE value less than or equal to that particular KGE value.

Evaluation of the performance for the different model versions provides insights into the impacts of the model calibrations and process representations. Comparison of E-HYPE (DK) versions v4.1.1 (Median KGE = 0.50) and v4.3.0 (Median KGE = 0.48) shows a similar shape for cumulative probabilities $\geq 50\%$ but diverging tails for probabilities $< 50\%$ (Fig. 4). This indicates that the new snowmelt routine and E-HYPE (DK) recalibration from version 4.1.1 to v4.3.0 helped to reduce the number of severely underperforming sites, although performance declined at several others (Figs. 3 and 4). Compared to E-HYPE (DK), the DK-HYPE model provided higher KGE values for nearly all stations; median KGEs for versions 0.1, 0.2, and 1.0 were 0.60, 0.68, and 0.81, respectively. The result of the initial calibration of the DK-HYPE model (i.e. from v0.1 to v0.2) was similar to that of the recalibration of the E-HYPE (DK) model (i.e. from v4.1.1 to v4.3.0); improvements in DK-HYPE model performance reduced the number of underperforming sites (cumulative probabilities $< 50\%$) with only minor improvements in model performance at the best-performing sites. The greatest improvement in DK-HYPE performance occurred between versions v0.2 and v1.0 when the model was recalibrated and the revised groundwater representation was implemented; with these changes, large improvements in model performance were achieved for the middle-performance sites (cumulative probabilities from 30 to 80%). Of all the models and versions tested, the DK-LSTM model provided the best model performance (Median KGE = 0.91).

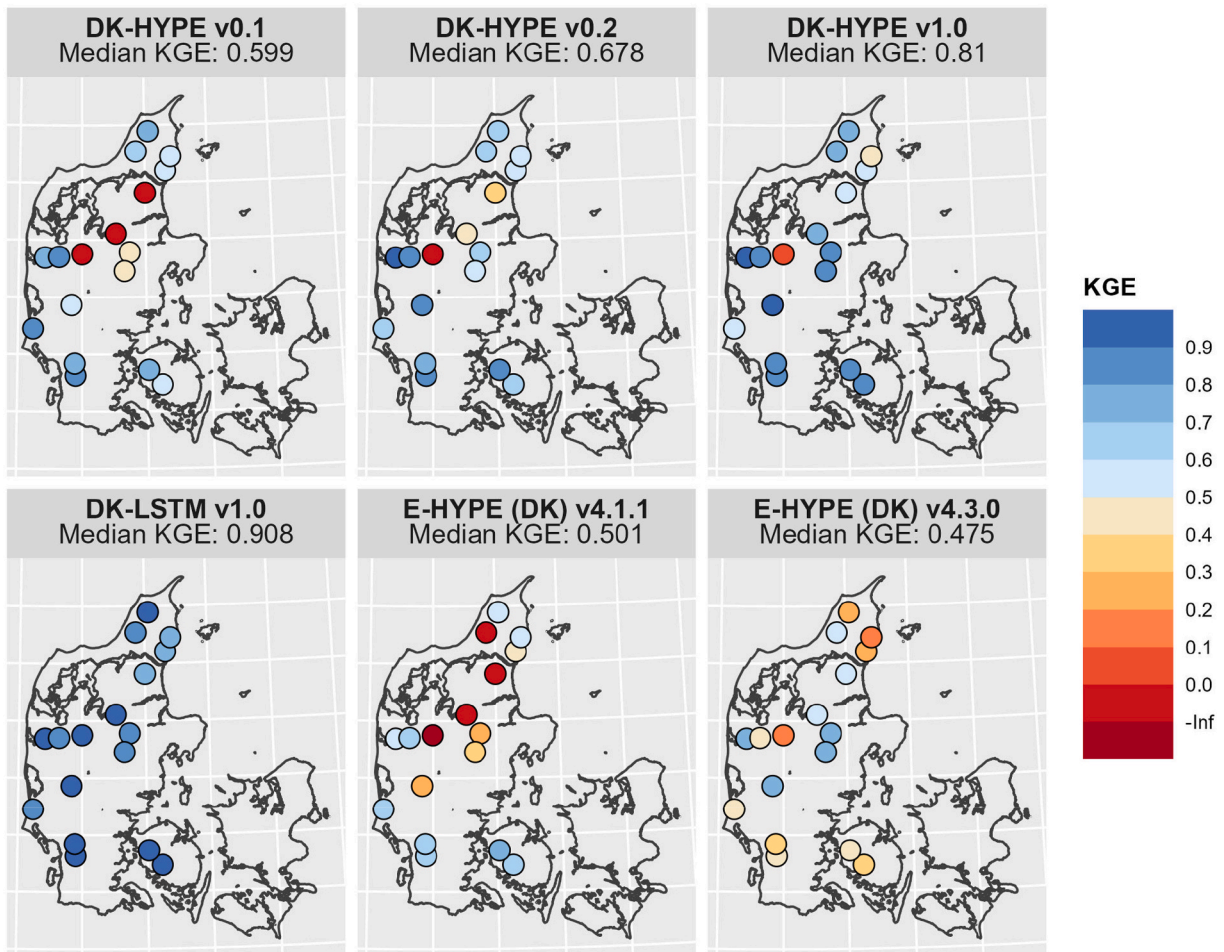


Fig. 3. Maps of model performance (KGE) by model version.

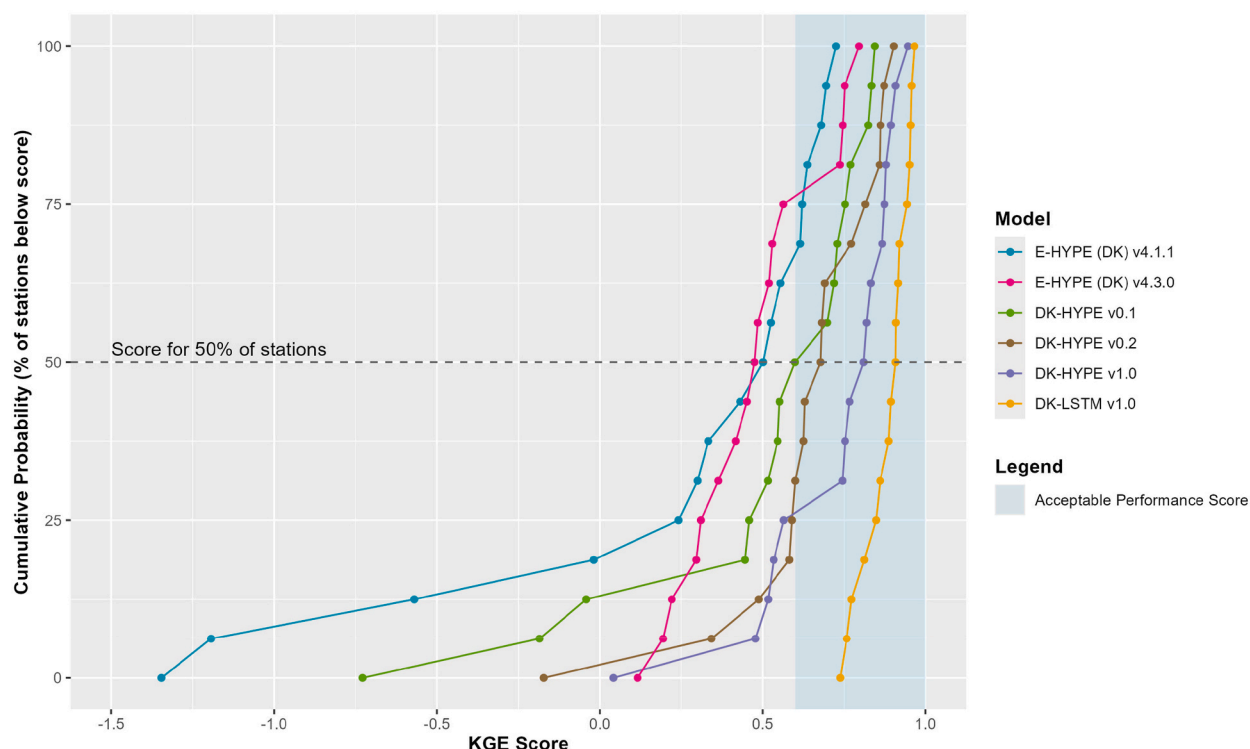


Fig. 4. Cumulative distribution function comparison of model performance (KGE) by model version.

The average peak-flow performance scores across the 17 stations for the period January 2023 to December 2024 (Table 4) show a similar trend as what can be observed from the KGE metric. The LSTM model performs best (average CSI = 0.32), and the E-HYPE (DK) models have the lowest scores (average CSI = 0.05 and 0.02 for versions 4.1.1 and 4.3.0 respectively). The DK-HYPE v1.0 model performs below the LSTM model (CSI = 0.25), with a lower hit rate but a higher success ratio. This indicates that the KGE is a reasonable metric to calibrate the flood warning model, and that the good performance of the LSTM model is not related to over-calibration.

3.2. Evaluation of hydrological responses during historical flood events

Simulations of significant flood events during December 2015 and February 2020 were used to evaluate the hydrological responses of the models/versions under different event conditions. Denmark received a significant amount of rain during November and December 2015, leading to widespread soil saturation and elevated river levels, and on December 25–26 heavy rainfall combined with the existing conditions resulted in a large runoff spike which flooded several key traffic hubs in Denmark [66]. In contrast, throughout February 2020, multiple large-scale frontal storm systems passed over Denmark in quick succession (notably on February 9th, 16th, and 21st), causing a compound flood event with sustained river discharge and significant flooding [67]. Timeseries of the simulated vs. observed discharges for the two events are presented in Fig. 5 for four observation stations (Storåen, Skjern Å, Ribe Å, and Gudenåen) that were included in each of the models. Discharge observations were not available for the Storåen station for the 2015 event. DMI's operational warning categories are defined by exceedance of specific return periods, with a Category 1 (Yellow) warning corresponding to a 5–20-year return period, a Category 2 (Orange) warning corresponding to a 20–50-year return period, and a Category 3 (Red) warning indicating return periods exceeding 50 years. In the operational system, the return periods are calculated on simulated time series from the operational model to account for local model biases. In Fig. 5, however, the return periods shown are calculated

Table 4

Average peak-flow performance metrics across the 17 stations for the period 2023 – 2024. CSI: Critical Success Index, HR: Hit Rate, SR: Success Ratio for the model's ability to predict flows exceeding a 2-year return period threshold.

Model Metric	DK-HYPE v0.1	DK-HYPE v0.2	DK-HYPE v1.0	E-HYPE (DK) v4.1.1	E-HYPE (DK) v4.3.0	DK-LSTM v1.0
CSI	0.11	0.17	0.25	0.05	0.02	0.32
HR	0.19	0.24	0.51	0.07	0.02	0.76
SR	0.33	0.43	0.49	0.20	0.33	0.40

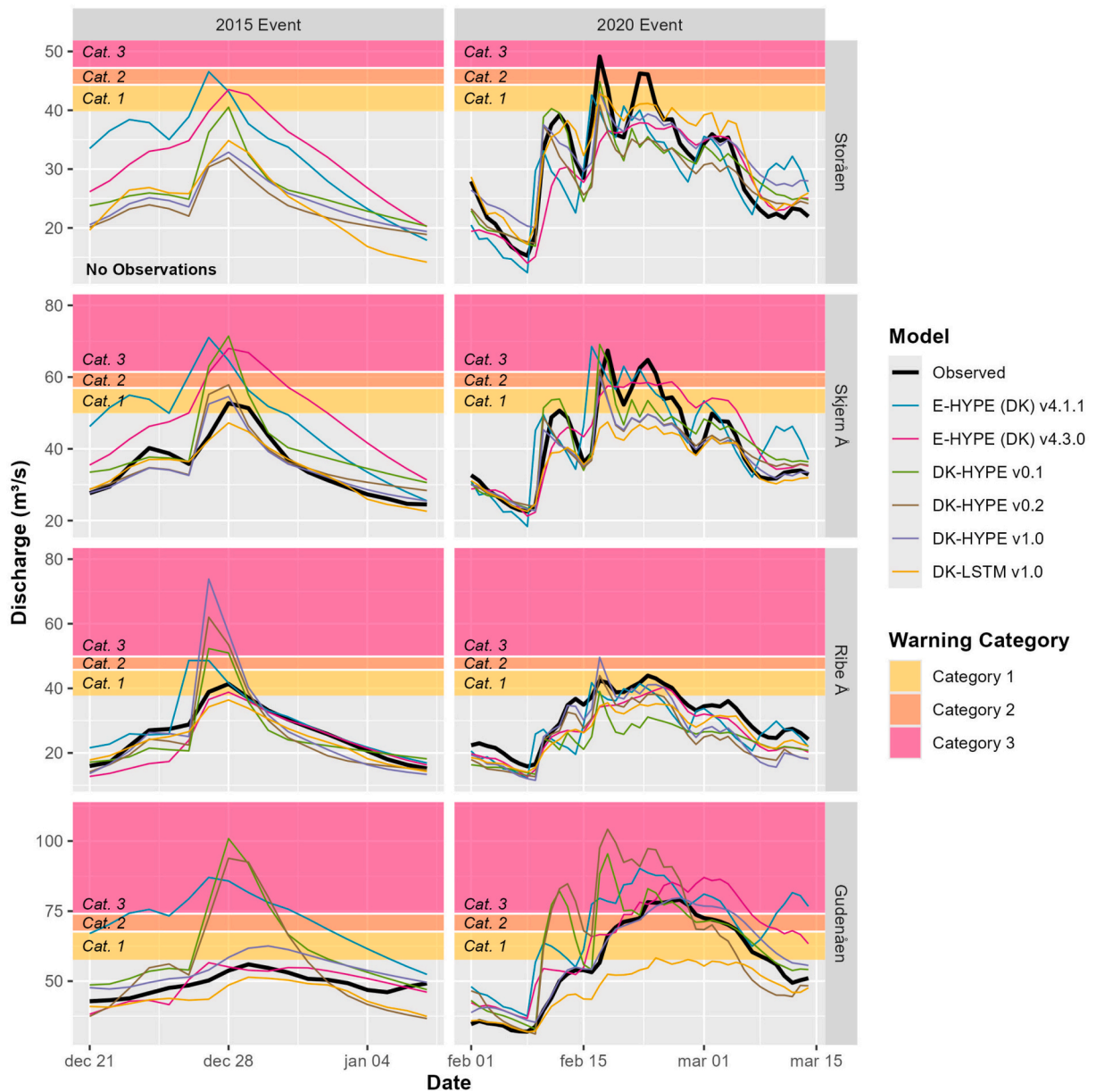


Fig. 5. Comparison of simulated vs. observed discharges at four stations (Storåen, Skjern Å, Ribe Å, and Gudenåen) for December 2015 and February 2020 flood events. Discharge observations were not available for the Storåen station for the 2015 event. Category 1 (Cat. 1) warnings correspond to a 5–20-year return period, Category 2 (Cat. 2) warnings correspond to a 20–50-year return period, and Category 3 (Cat. 3) warnings indicate return periods exceeding 50 years.

using the observed time series to provide a fairer model comparison.

Inspection of the December 2015 event reveals how the models respond to significant rainfall on already saturated catchments. Of the process-based models, the DK-HYPE model versions predicted sharper, flashier responses to the precipitation whereas the responses for the E-HYPE (DK) model versions were longer and smoother. The dampened response of the E-HYPE (DK) model may reflect the spatial averaging of the precipitation forcing data across the larger model catchments relative to the DK-HYPE model. The DK-LSTM model was able to closely match the shape of the observed hydrographs but tended to underestimate the magnitude of the peaks. Observed peak flows at the Skjern Å and Ribe Å stations culminated within the Category 1 warning threshold whereas Gudenåen peaked just below it. Overall, the models tended to overestimate the warning thresholds, with E-HYPE v4.1.1, DK-HYPE v0.1, and DK-HYPE v0.2 predicting Category 2–3 peaks at all three stations. DK-HYPE v1.0 and DK-LSTM each correctly predicted the correct warning category for one station but showed opposing biases with DK-HYPE v1.0 overestimating and DK-LSTM underestimating the warning categories for the other stations. E-HYPE v4.3.0 performed the best overall with correct predictions at Ribe Å and Gudenåen

but overprediction of the warning category for Skjern Å.

Analysis of the February 2020 events shows how the models respond to a period of persistent, overlapping weather systems. Of the three models, the DK-HYPE model versions captured the inter-event dynamics best and provided the most clearly defined flood peaks. The E-HYPE (DK) v4.1.1 simulation also captured the general shape of the different flood peaks, but the peaks of the E-HYPE (DK) v4.3.0 simulation were flattened and dampened, potentially indicating a bias in the model towards partitioning water into storage components rather than runoff pathways. Compared to the process-based models, the DK-LSTM model displayed an intermediate response between the fast-responding DK-HYPE model and the more dampened E-HYPE (DK) model. Across the Storåen, Skjern Å, and Ribe Å sites, all three models tended to underestimate the flood peaks, leading to underprediction of both warning threshold exceedance and flood severity across successive events (especially for the third event in the series on February 21st). At the Gudenåen station, however, the storm peaks merged into a single large peak that was overestimated by the HYPE models and underestimated by the LSTM model. These results suggest that the models only partially capture the cumulative and amplifying effects of compound storm events.

3.3. Model performance vs. development time

After the end of the project, an estimate of the time spent on each model version was compiled (Section 2.3). Fig. 6 presents a comparison between the model performance for each model/version and the estimated time spent on its development. In this figure, the values along the y-axis indicate the KGE values for the 17 comparison stations (Fig. 2), and the width of the boxplots represents the estimated cumulative development time for each model (Table 2). These time estimates are presented as ranges to highlight the uncertainty in the estimates since the time spent on model development was not explicitly tracked during the project and because people were not working full-time on the project. Furthermore, the time estimates in Fig. 6 are made with the assumption that the large-scale E-HYPE model was already constructed, as was the case in Denmark (Fig. 1). Thus, the E-HYPE (DK) v4.1.1 model setup has essentially zero development time and the only time investment in terms of the project was one day to extract the Danish model domain from the pan-European E-HYPE model setup and to select a parameter set from the existing parameter ensemble (Section 2.2.1). The time estimates for successive model versions (e.g. from DK-HYPE v0.1 to v0.2 to v1.0) include the time invested in the

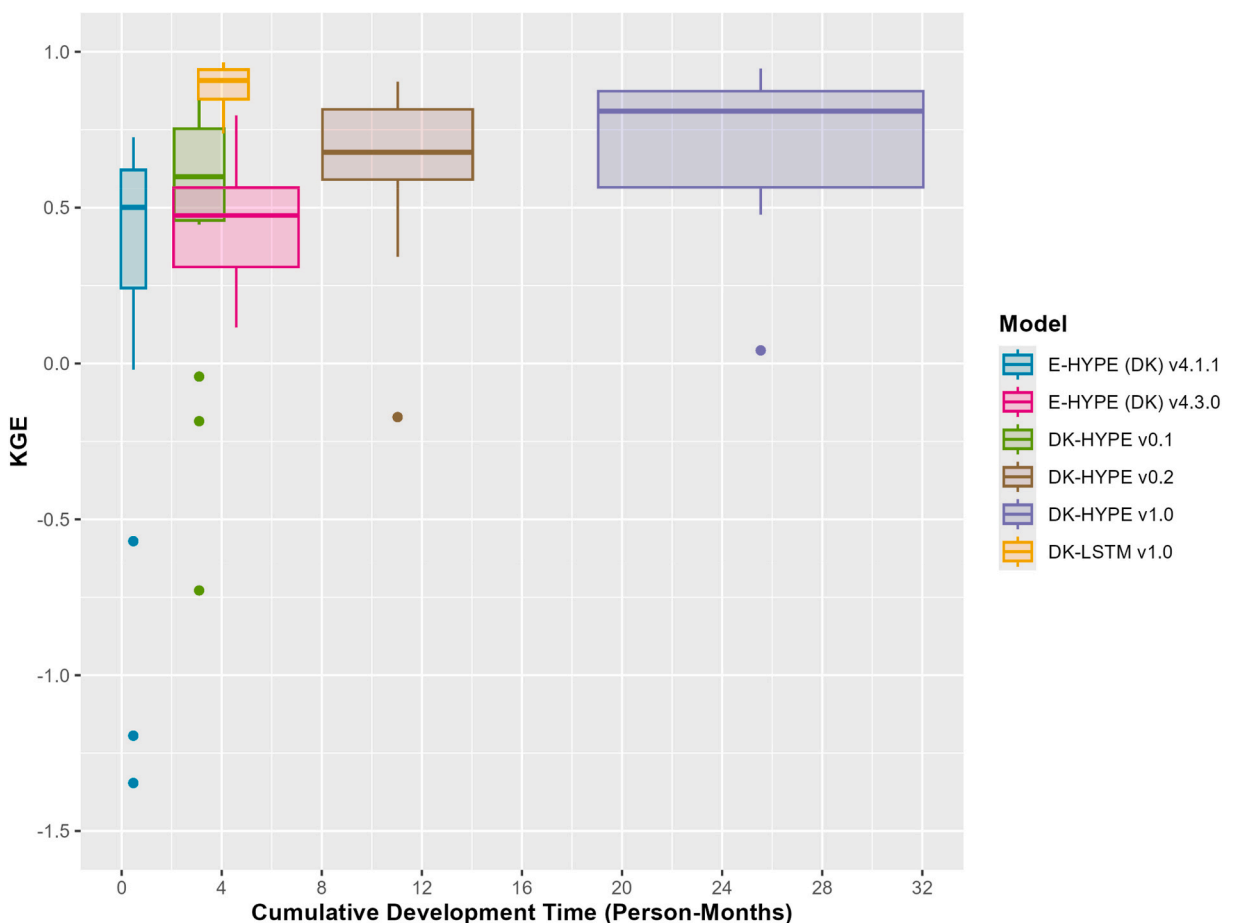


Fig. 6. Model performance vs. cumulative development time. Width of the boxplots represents the uncertainty in the time estimates.

previous model versions. Overall, the model with the shortest time until operational readiness was the top-down process-based E-HYPE (DK) v4.1.1 setup followed by the bottom-up process-based DK-HYPE v0.1 setup, the bottom-up data-driven DK-LSTM model, E-HYPE (DK) v4.3.0, and DK-HYPE v0.2 and v1.0 (Table 2).

Examination of the relationship between model performance and time investment provides insights into the efficiency of model development for the different modeling approaches and model types. Inspection of Fig. 6 shows that the additional time spent recalibrating the E-HYPE (DK) model between version v4.1.1 and v4.3.0 reduced the spread of the KGE values, but that average model performance did not improve significantly. Furthermore, the median KGE of E-HYPE (DK) v4.3.0 ($KGE = 0.48$) was less than that of the uncalibrated DK-HYPE v0.1 ($KGE = 0.60$), which suggests that the model performance for the top-down E-HYPE (DK) model was limited by the forcing data and model structure (e.g. delineation, routing, groundwater representation, etc.) rather than by the model parameterization. Compared to the top-down E-HYPE (DK) model, the bottom-up DK-HYPE model provided better model performance but required a longer development time. While the initial calibration of the DK-HYPE model from version v0.1 to v0.2 led to slight improvements in model performance (from median $KGE = 0.60$ to 0.68), the greatest improvement in model performance corresponded to the implementation of the revised groundwater representation from version v0.2 to v1.0 (from median $KGE = 0.68$ to 0.81). This also suggests that, like the E-HYPE (DK) model, the DK-HYPE model performance is influenced more by the process representation than by the model parameterization. Compared to the process-based HYPE models, the DK-LSTM model provided the best model performance (Median $KGE = 0.91$) and required a relatively short development time (3-5 person-months) until operational readiness.

4. Discussion

4.1. When do different modeling approaches offer the greatest benefit?

Evaluation of the top-down vs. bottom-up process-based models showed a trade-off between model performance and time to operational readiness. Overall, the top-down E-HYPE (DK) model provided the greatest value early in the project (Figs. 1 and 6) by providing nearly immediate access to model forecasts and insights into how a HYPE-based forecasting chain could be constructed. This significantly expedited the development of the operational flood forecast system by allowing DMI to design the forecast chain and warning system interface in parallel with the bottom-up DK-HYPE model (i.e. instead of developing the model first and the rest of the system afterwards). However, the long-term usefulness of the E-HYPE (DK) model was limited by its model performance, which could not be improved meaningfully despite additional calibration efforts between E-HYPE (DK) v4.1.1 and v4.3.0 (Fig. 6). Revising the E-HYPE (DK) model representation to include processes such as groundwater could have potentially improved model performance, but this would have necessitated additional development time, and the model would likely still be constrained by the quality and resolution of the other input data. In contrast, the value of the bottom-up DK-HYPE model was greatest later in the project as it offered better overall model performance and confidence in flood forecasts. Furthermore, the DK-HYPE model allowed for the inclusion of national data not available from European or global datasets to achieve more localized forecasts and higher spatial resolution forecasting capabilities. However, these benefits came at the cost of a significantly longer development period (Figs. 1 and 6).

Compared to the process-based models, the data-driven model offered a higher return in model performance for time invested in development. After four person-months of development time, the bottom-up, data-driven DK-LSTM model had a median KGE of 0.91 across the comparison stations. In contrast, the performance of the two process-based models was much lower after a similar time investment; the median KGE for DK-HYPE v0.1 was 0.60 after 3 person-months, and the median KGE for E-HYPE (DK) v4.3.0 was 0.48 after 4.5 person-months (Fig. 6). Although development times could differ if the LSTM model was built with different architecture or if the HYPE models were calibrated for different parameters, this demonstrates that the data-driven model provides the most effective short-term performance. Similarly, Liu et al. [68] show an increased hindcast KGE performance of a national LSTM model (Mean $KGE = 0.76$) for Denmark compared to a physics-based national hydrological model (mean $KGE = 0.56$). The initial model development work within this project focused on the process-based models, so development of the DK-LSTM model did not commence until nearly two years after the start of the project (Fig. 1). Reflecting on this, if no large-scale process-based model (e.g. E-HYPE) is available to provide an initial model setup during early system development, then an LSTM model could serve as a viable alternative due to its rapid development and strong predictive performance. However, the strong performance of DK-LSTM should be viewed with due consideration to the different calibration setups: DK-LSTM was trained with data from all comparison stations with temporal hold-out periods, whereas the HYPE models were calibrated on a subset of the comparison stations.

Together, these results highlight the complementary strengths of different modeling approaches. With unconstrained time and resources, a bottom-up process-based approach is likely to yield the most favorable balance between explainability and forecasting performance. In contrast, for rapid development of flood modeling or warning systems, it may be preferable to use a data-driven or top-down process-based approach. Ultimately, the hybrid strategy taken in this study (starting with a top-down model followed by bottom-up development) proved effective in balancing short-term efficiency with long-term quality.

4.2. What levels of model performance and physical explainability are sufficient?

An important consideration when selecting a modeling approach is what level of model performance is sufficient. The answer to this question is inherently dependent on the time and resources available, as expectations after a short development period (e.g. one month) are naturally lower than those after a year of development work. A formal evaluation of forecast skill was outside the scope of this paper, but analysis of the simulated hydrological responses for the two historical flood events (Section 3.2, Fig. 5) showed that the E-HYPE (DK) model performed better than the DK-HYPE model in predicting flood peak magnitude and warning category, despite its

lower KGE values. The E-HYPE (DK) model may therefore provide sufficient performance for operational decision making, particularly in the early stages of a system, until bottom-up or data-driven approaches have been developed. These findings highlight the potential value of top-down modeling approaches for data- and/or resource-limited regions, where refining a top-down model may be more feasible than developing a bottom-up alternative, while still delivering sufficient forecasting value. For example, in the experience of national hydrological services in West Africa, the accuracy of forecasted warning categories was higher than the accuracy of forecasted discharge from a HYPE model adapted for the region using a top-down approach [7]. Moreover, despite varying KGE values, the forecasted warning categories were accurate enough to be used operationally to reduce infrastructure damage and save lives [69].

Beyond the initial goal of creating a functional model, it is also important to establish a solid physical foundation to support continued model development and to provide evidence for warning decisions. Thus, priority should be given to developing a physically consistent and hydrologically plausible model rather than one which manages to match the actual hydrograph but looks unreasonable when one looks at the underlying water balance; even if a hydrological model cannot precisely predict the exact shape and timing of a flood peak, it may still provide sufficient information to inform operational flood warning decisions. In comparison to the process-based models, the data-driven DK-LSTM model provided the highest model performance while also requiring less development time. However, we have observed that the DK-LSTM model fails to make reasonable predictions during black swan events (extremely rare, unpredictable events with widespread impacts; See Supplemental Information Section S3), in line with recent studies ([70,71 [72]]). Furthermore, in a forecasting context, the internal states in process-based models (e.g. runoff, soil moisture, and snow water equivalent) can assist the forecasters in making informed decisions. In data-driven models, these states are not directly interpretable as they are encoded in a latent representation. Although not available within the DK-LSTM model architecture, LSTM models can indeed internalize hydrological concepts ([37][73]) and internal hydrological states can be unveiled using probing strategies [74]. In summary, process-based models with lower model-fit metrics should not be discounted solely due to performance, as their physical basis and internal states provide inherent value. Data-driven models, while offering potentially higher predictive performance within their training domain, require caution when applied beyond it.

4.3. What institutional and practical lessons can be drawn from the development process?

DMI's success in building an operational forecasting system within the deadline can be partially attributed to several key decisions, namely investing in both process-based and data-driven models, partnering with expert organizations, using open-access tools, and performing pilot testing. Based on DMI's experience, those aiming to build flood warning systems from scratch will benefit greatly by investing in a combination of process-based and data-driven models, especially for forecasting unprecedented extremes. Moreover, institutions can also achieve greater success and quick turnarounds by strategically investing in similar partnerships with flagship modeling organizations and universities to learn the best of existing methods and state-of-the-art practices. Rapid in-house model development was also aided by DMI's choice to use community, open access tools that were freely available and quick to deploy. Furthermore, pilot testing of the HYPE and LSTM modeling software through small pilot projects (Fig. 1) provided the development team with valuable experience which allowed them to be more efficient in setting up DK-HYPE and DK-LSTM. We further note that model development efficiency tends to improve over time as staff gain experience with available datasets and model software, data quality progressively improves when errors are identified, and tools for data processing accumulate and mature. Such pilots could give other emerging authorities a solid baseline to assess the resources their teams may need for developing full-scale models. Collectively, these decisions enabled the timely delivery of a reliable and operational forecasting system.

4.3.1. Top-down process-based modeling

The critical consideration when pursuing a top-down modeling approach is assessing whether its performance is adequate for the target application. Extracting a domain from a large-scale model is straightforward due to the prevalence of continental and global models such as E-HYPE [30] and World-Wide HYPE [75]. However, model performance may vary depending on location and dominant hydrological processes [30]. In Denmark, the hydrological regime is largely groundwater dominated and many coastal catchments are flashy. Storm rainfall often occurs as well-defined frontal systems or summer convection, causing precipitation during events to vary significantly over short distances. These dynamics are thus better captured by a higher resolution locally calibrated model than a large-scale model. For locations where local hydrology is comprised of large transboundary basins or areas with more homogenous hydro-climatic conditions, the top-down effort may yield better performance. However, even if it may be desirable to develop a bottom-up model, significant data and resources are required to support its development. The results showed that the top-down E-HYPE (DK) model performed better than the bottom-up DK-HYPE model in predicting flood peak magnitude and warning category (Section 3.2) and could therefore provide adequate performance for operational decision making despite its lower KGE values. Thus, when sufficient data or resources are not available to support a bottom-up model, it may be more strategic to adopt a top-down approach and focus on developing the operational forecasting pipeline, from data acquisition to a final visualization portal, while deferring bottom-up model development until data and resources become available.

4.3.2. Bottom-up process-based modeling

The key goal for the development of a bottom-up process-based modeling approach should ultimately be to exploit local data sources and improve process representation not available within a top-down approach. Additional calibration of the top-down E-HYPE (DK) model produced only marginal performance improvements (Fig. 6) which suggests that the top-down model performance is limited by the model structure (e.g. subbasin delineation and process representations) and input data rather than the model parameterization. Thus, further improvements to the top-down model performance cannot be obtained without revising the model structure.

In contrast, the data sources used to construct the bottom-up process-based DK-HYPE model allowed for a higher resolution model that could be parameterized to local hydro-climatological conditions (e.g. the dominant role of groundwater) and improve the model forecast performance beyond that of the top-down model.

4.3.3. Data-driven modeling

Lastly, DMI's development of the DK-LSTM was opportunity-driven and motivated by the growing focus in hydrological research on the potential of this approach. Our experience here shows that, provided sufficient high-quality data are available, LSTM model development can be carried out very rapidly with good overall performance. The availability of longer time series of dynamic data, such as meteorological forcings and stream discharge, is especially important in achieving good model performance in gauged catchments. In our case, sufficient dynamic data was 10+ years of continuous meteorological forcing data and stream discharge observations with a minimum of 3 years scattered over the full data coverage timespan. For data scarce regions, transfer learning from other regions could be considered, but this will in return require more development time. The drawbacks of the LSTM model, such as its limited ability to reproduce extreme events, underline the need for a model ensemble approach, where data-driven models are supported by some level of process-based decision support. Ultimately, given the rapid pace of development in the flood modeling community, emerging forecasting authorities should actively engage and invest in ongoing research, while continuously integrating in-house model development with the latest scientific advances.

5. Conclusions

This study provides practical guidance on how to select a hydrological modeling pathway for developing an operational flood forecasting system. It does so through a case study, where a newly appointed flood warning authority had to develop a national system from scratch under a strict deadline of three years. The study presents a unique quantification of performance, development time, and operational readiness of three model development strategies: (1) "top-down" which uses a pan-European, process-based model cropped to Denmark with coarse forcing data, (2) "bottom-up" where a process-based model is developed from local data, and (3) "data-driven" where an LSTM model is developed from local data.

The results highlighted the tradeoffs between the different modeling approaches. The top-down process-based approach was extremely efficient in terms of development time and provided the fastest pathway to operational deployment. However, model performance was constrained and did not significantly improve through additional calibration attempts. Despite this, the top-down process-based models performed comparably to their bottom-up counterparts in predicting flood peak magnitude for two historical events, highlighting their suitability for rapid implementation of operational systems and for use in areas lacking quality local data. The bottom-up process-based approach leveraged detailed local geographic and forcing data to improve process representation and overall model performance but required longer development time and local expert knowledge. Since this study is a retrospective analysis of operational model development and not a pre-designed scientific experiment, there was no common set of validation stations held out for thorough model validation, and the results should be considered with this limitation in mind. Finally, although the data-driven model had the highest overall performance metrics, it underestimated the observed flood peaks during the two historical events and produced unrealistically low predictions when forced with extreme amounts of rainfall (See Supplemental Information Section S3). Furthermore, process-based models provide physical interpretability lacking in data-driven models, which can assist forecasters in making informed operational decisions. Thus, models should not be discounted for operational use solely on the basis of their model-fit metrics.

While no universally optimal model development strategy exists, the selection of a modeling pathway should be guided by constraints on time, budget, available data, and operational objectives. High-quality local data were a key factor enabling both bottom-up models (process-based and data-driven) to achieve performance beyond that of the top-down approach. Institutionally, partnering with experienced organizations, using open-source tools, and performing local pilot studies proved valuable in building internal expertise. In addition, the hybrid approach of pursuing multiple modeling approaches in parallel proved effective in balancing short-term operational efficiency with long-term model performance. We hope that the practical guidance we provide here can support other emerging flood warning authorities in establishing forecasting systems, assessing the resources needed to develop and operationalize them, and ultimately bring us one step closer to delivering flood early warnings for all. Likely or not, it's a dream that we keep, and if this is a sign, then it's only a matter of time.

CRedit authorship contribution statement

Conrad Brendel: Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Raphaël Payet-Burin:** Methodology, Formal analysis. **Phillip D. Aarestrup:** Methodology, Formal analysis. **Sanita Dhaubanjari:** Writing – original draft, Investigation, Formal analysis, Data curation. **Maggie H. Madsen:** Methodology, Formal analysis. **Grith Martinsen:** Writing – original draft, Methodology, Formal analysis, Data curation. **Lucas Dalgaard Jensen:** Methodology, Formal analysis. **Jonas W. Pedersen:** Writing – original draft, Methodology, Formal analysis. **Charlotte A. Plum:** Methodology, Formal analysis. **Emma D. Thomassen:** Writing – original draft, Methodology, Formal analysis. **Cecilie Thrysøe:** Methodology, Formal analysis. **René Capell:** Methodology, Conceptualization. **Jafet C.M. Andersson:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization. **Michael B. Butts:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work the authors used ChatGPT and Copilot in order to improve language and readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding sources

The collaboration between DMI and SMHI was made possible by support from the Danish government under the Finance Act 2022.

Declaration of competing interests

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.

Acknowledgements

With the Finance Act for 2022, DMI was given the responsibility for flood warning in Denmark and the development of a national flood forecasting and early warning system. This work has been carried out in collaboration with the Danish Agency for Green Transition and the Aquatic Environment (SGAV), the Geological Survey of Denmark and Greenland (GEUS), the Danish Agency for Climate Data (KDS), the Danish Environmental Portal (DMP) and the Danish Coastal Authority (KDI). In particular, the authors wish to acknowledge the support and contributions of Special Consultant Rasmus Savange Ringgaard and his colleagues in investigating and analyzing the Danish river data. The authors wish to acknowledge the support of Vejle Municipality, in particular, Engineer Paul Landsfeldt and Climate Coordinator Ulla Pia Geertsen who have been extremely helpful in providing both data and information for the pilot study in Vejle. The collaboration between DMI and SMHI was made possible by support from the Danish government under the Finance Act 2022. Finally, special thanks to Dream Theater for the inspiration for the title of this article \m/

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ijdr.2026.106393>.

Data availability

Data will be made available on request.

References

- [1] UNDRR, *The Human Cost of Disasters: an Overview of the Last 20 Years (2000–2019)*, United Nations Office for Disaster Risk Reduction, Geneva, 2020.
- [2] UNDRR, *Global Assessment Report on Disaster Risk Reduction 2025: Resilience Pays: Financing and Investing for our Future*, Stylus Publishing, LLC, 2025.
- [3] WMO, UNDRR, *Global Status of Multi-Hazard Early Warning Systems: 2025, 2025*.
- [4] L. Alfieri, P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, F. Pappenberger, GloFAS - global ensemble streamflow forecasting and flood early warning, *Hydrol. Earth Syst. Sci.* 17 (2013) 1161–1175, <https://doi.org/10.5194/hess-17-1161-2013>.
- [5] G. Nearing, D. Cohen, V. Dube, M. Gauch, O. Gilon, S. Harrigan, A. Hassidim, D. Klotz, F. Kratzert, A. Metzger, S. Nevo, F. Pappenberger, C. Prudhomme, G. Shalev, S. Shenzen, T.Y. Tekalign, D. Weitzner, Y. Matias, Global prediction of extreme floods in ungauged watersheds, *Nature* 627 (2024) 559–563, <https://doi.org/10.1038/s41586-024-07145-1>.
- [6] J. Thielen, J. Bartholmes, M.-H. Ramos, A. De Roo, The European flood alert system—part 1: concept and development, *Hydrol. Earth Syst. Sci.* 13 (2009) 125–140.
- [7] J. Lienert, J.C.M. Andersson, D. Hofmann, F. Silva Pinto, M. Kuller, The role of multi-criteria decision analysis in a transdisciplinary process: co-developing a flood forecasting system in western Africa, *Hydrol. Earth Syst. Sci.* 26 (2022) 2899–2922, <https://doi.org/10.5194/hess-26-2899-2022>.
- [8] T.E. Adams III, *Flood forecasting in the United States NOAA/national weather service*, in: *Flood Forecasting*, Elsevier, 2016, pp. 249–310.
- [9] W.H. Azad, M.H. Hassan, N.H.M. Ghazali, A. Weisgerber, F. Ahmad, National flood forecasting and warning system of Malaysia: an overview, in: *International Conference on Dam Safety Management and Engineering*, Springer, 2019, pp. 264–273.
- [10] M. Girons Lopez, L. Crochemore, I.G. Pechlivanidis, Benchmarking an operational hydrological model for providing seasonal forecasts in Sweden, *Hydrol. Earth Syst. Sci.* 25 (2021) 1189–1209, <https://doi.org/10.5194/hess-25-1189-2021>.
- [11] W.F. Krajewski, D. Ceynar, I. Demir, R. Goska, A. Kruger, C. Langel, R. Mantilla, J. Niemeier, F. Quintero, B.-C. Seo, S.J. Small, L.J. Weber, N.C. Young, Real-time flood forecasting and information system for the state of Iowa, *Bull. Am. Meteorol. Soc.* 98 (2017) 539–554, <https://doi.org/10.1175/BAMS-D-15-00243.1>.
- [12] H. Kreibich, M. Müller, K. Schröter, A.H. Thieken, New insights into flood warning reception and emergency response by affected parties, *Nat. Hazards Earth Syst. Sci.* 17 (2017) 2075–2092, <https://doi.org/10.5194/nhess-17-2075-2017>.
- [13] E. Stephens, H. Cloke, Improving flood forecasts for better flood preparedness in the UK (and beyond), *Geogr. J.* 180 (2014) 310–316.
- [14] I. Fraga, L. Cea, J. Puertas, MERLIN: a flood hazard forecasting system for coastal river reaches, *Nat. Hazards* 100 (2020) 1171–1193, <https://doi.org/10.1007/s11069-020-03855-7>.
- [15] D. Fernández-Nóvoa, O. García-Feal, J. González-Cao, C. de Gonzalo, J.A. Rodríguez-Suárez, C. Ruiz del Portal, M. Gómez-Gesteira, MIDAS: a new integrated flood early warning system for the miño river, *Water* 12 (2020) 2319, <https://doi.org/10.3390/w12092319>.
- [16] G. Massazza, V. Tarchiani, J.C.M. Andersson, A. Ali, M.H. Ibrahim, A. Pezzoli, T. De Filippis, L. Rocchi, B. Minoungou, D. Gustafsson, M. Rosso, Downscaling regional hydrological forecast for operational use in local early warning: HYPE models in the Sirba River, *Water* 12 (2020) 3504, <https://doi.org/10.3390/w12123504>.

- [17] C.E. Brendel, R.L. Dymond, M.F. Aguilar, Integration of quantitative precipitation forecasts with real-time hydrology and hydraulics modeling towards probabilistic forecasting of urban flooding, *Environ. Model. Software* 134 (2020) 104864, <https://doi.org/10.1016/j.envsoft.2020.104864>.
- [18] M. Staudinger, A. Herzog, R. Loritz, T. Houska, S. Pool, D. Spieler, P.D. Wagner, J. Mai, J. Kiesel, S. Thober, B. Guse, U. Ehret, How well do process-based and data-driven hydrological models learn from limited discharge data? *Hydrol. Earth Syst. Sci.* 29 (2025) 5005–5029, <https://doi.org/10.5194/hess-29-5005-2025>.
- [19] J. Wang, H. Yousefi, M. Shourian, Comparison of process-based hydrological modeling and deep learning approaches for streamflow simulation, *Water Resour. Manag.* (2025), <https://doi.org/10.1007/s11269-025-04313-4>.
- [20] Y. Zhang, S. Ragetti, P. Molnar, O. Fink, N. Peleg, Generalization of an encoder-decoder LSTM model for flood prediction in ungauged catchments, *J. Hydrol.* 614 (2022) 128577, <https://doi.org/10.1016/j.jhydrol.2022.128577>.
- [21] G. Özdoğan-Sarıkoç, F. Dadaser-Celik, Physically based vs. data-driven models for streamflow and reservoir volume prediction at a data-scarce semi-arid basin, *Environ. Sci. Pollut. Res.* 31 (2024) 39098–39119, <https://doi.org/10.1007/s11356-024-33732-w>.
- [22] Z. Guo, V. Moosavi, J.P. Leitão, Data-driven rapid flood prediction mapping with catchment generalizability, *J. Hydrol.* 609 (2022) 127726.
- [23] F.J. Hop, R. Linneman, B. Schnitzler, A. Bomers, M.J. Booij, Real time probabilistic inundation forecasts using a LSTM neural network, *J. Hydrol.* 635 (2024) 131082, <https://doi.org/10.1016/j.jhydrol.2024.131082>.
- [24] S. Nevo, E. Morin, A. Gerzi Rosenthal, A. Metzger, C. Barshai, D. Weitzner, D. Voloshin, F. Kratzert, G. Elidan, G. Dror, G. Begelman, G. Nearing, G. Shalev, H. Noga, I. Shavitt, L. Yuklea, M. Royz, N. Giladi, N. Peled Levi, O. Reich, O. Gilon, R. Maor, S. Timnat, T. Shechter, V. Anisimov, Y. Gigi, Y. Levin, Z. Moshe, Z. Ben-Haim, A. Hassidim, Y. Matias, Flood forecasting with machine learning models in an operational framework, *Hydrol. Earth Syst. Sci.* 26 (2022) 4013–4032, <https://doi.org/10.5194/hess-26-4013-2022>.
- [25] Y. Xu, K. Lin, C. Hu, S. Wang, Q. Wu, L. Zhang, G. Ran, Deep transfer learning based on transformer for flood forecasting in data-sparse basins, *J. Hydrol.* 625 (2023) 129956, <https://doi.org/10.1016/j.jhydrol.2023.129956>.
- [26] L.B.L. Santos, E.V. Escobar-Silva, L.F. Satolo, R.S. Oyarzabal, M.M. Diniz, R.G. Negri, G.R.T. Lima, S. Stephany, J.A.J.P. Soares, J.S. Duque, F.L. Saraiva Filho, A. N. Caseri, L. Bacelar, Machine learning-based hydrological models for flash floods: a systematic literature review, *Smart Constr. Sustain. Cities* 3 (2025) 21, <https://doi.org/10.1007/s44268-025-00071-9>.
- [27] M.A.D. Larsen, G. Karamitilios, M.L. Dømgård, K. Halsnaes, DMI Report 21-35 Data Driven Climate Change Adaptation Part B: National and Local Scale Flood Modelling as a Basis for Damage Cost Assessments, 2021.
- [28] P. Skougaard Kaspersen, N. Hoegh Ravn, K. Arnbjerg-Nielsen, H. Madsen, M. Drews, Comparison of the impacts of urban development and climate change on exposing European cities to pluvial flooding, *Hydrol. Earth Syst. Sci.* 21 (2017) 4131–4147, <https://doi.org/10.5194/hess-21-4131-2017>.
- [29] Q. Zhou, P.S. Mikkelsen, K. Halsnaes, K. Arnbjerg-Nielsen, Framework for economic pluvial flood risk assessment considering climate change effects and adaptation benefits, *J. Hydrol.* 414–415 (2012) 539–549, <https://doi.org/10.1016/j.jhydrol.2011.11.031>.
- [30] C. Brendel, R. Capell, A. Bartosova, 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, *J. Hydrol.: Reg. Stud.* 50 (2023) 101544, <https://doi.org/10.1016/j.ejrh.2023.101544>.
- [31] K.P. Georgakakos, D.-J. Seo, H. Gupta, J. Schaaek, M.B. Butts, Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *J. Hydrol., The Distributed Model Intercomparison Project (DMIP)* 298 (2004) 222–241, <https://doi.org/10.1016/j.jhydrol.2004.03.037>.
- [32] T. Roy, A. Serrat-Capdevila, H. Gupta, J. Valdes, A platform for probabilistic multimodel and multiproduct streamflow forecasting, *Water Resour. Res.* 53 (2017) 376–399, <https://doi.org/10.1002/2016WR019752>.
- [33] C. Thébaud, W.J.M. Knoben, N. Addor, A.J. Newman, D. Spieler, N.A. Vásquez, Y. Song, G.J. Gründemann, S. Carney, M. Kumar, K. van Werkhoven, C. Shen, A. W. Wood, M.P. Clark, Comparing multi-model mosaic and multi-model combination methods to simulate streamflow across the contiguous USA, *EGU sphere* (2026) 1–43, <https://doi.org/10.5194/egusphere-2025-6083>.
- [34] SMHI, HYPER model documentation. <http://hype.smhi.net/wiki/doku.php>, 2025.
- [35] S. Hochreiter, J. Schmidhuber, Long Short-Term memory, *Neural Comput.* 9 (1997) 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>.
- [36] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, M. Herrnegger, Rainfall–runoff modelling using long short-term memory (LSTM) networks, *Hydrol. Earth Syst. Sci.* 22 (2018) 6005–6022, <https://doi.org/10.5194/hess-22-6005-2018>.
- [37] F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, G. Nearing, Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, *Hydrol. Earth Syst. Sci.* 23 (12) (2019) 5089–5110, <https://doi.org/10.5194/hess-23-5089-2019>.
- [38] D. Feng, K. Fang, C. Shen, Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales, *Water Resour. Res.* 56 (2020), <https://doi.org/10.1029/2019WR026793> e2019WR026793.
- [39] J. Koch, R. Schneider, Long short-term memory networks enhance rainfall-runoff modelling at the national scale of Denmark, *GEUS Bull.* 49 (2022), <https://doi.org/10.34194/geusb.v49.8292>.
- [40] D. Yamazaki, D. Ikeshima, J. Sosa, P.D. Bates, G.H. Allen, T.M. Pavelsky, MERIT hydro: a high-resolution global hydrography map based on latest topography dataset, *Water Resour. Res.* 55 (2019) 5053–5073, <https://doi.org/10.1029/2019WR024873>.
- [41] European Union, Copernicus Land Monitoring Service, EU-Hydro – River network database, version 1.3. <https://land.copernicus.eu/imagery-in-situ/eu-hydro/eu-hydro-river-network-database>, 2019.
- [42] P. Berg, F. Almén, D. Bozhinova, 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 (2021) 1531–1545, <https://doi.org/10.5194/essd-13-1531-2021>.
- [43] GRDC, The global runoff data centre, 56068 koblenz, Germany. https://grdc.bafg.de/data/data_portal/, 2026.
- [44] European Union, Copernicus Land Monitoring Service, Corine land cover (CLC) 2006, version 2020_20u1. <https://land.copernicus.eu/pan-european/corine-land-cover/clc-2006>, 2020.
- [45] ESA, CCI land cover dataset (v 1.6.1). <http://maps.elie.ucl.ac.be/CCI/viewer/download.php>, 2016.
- [46] RGI Consortium, Randolph glacier inventory - a dataset of global glacier outlines, version 6. Natl. Snow ice data cent. <https://nsidc.org/data/nsidc-0770/versions/6>, 2017.
- [47] M. Kottek, J. Grieser, C. Beck, B. Rudolf, F. Rubel, World map of the Köppen-Geiger climate classification updated, *Meteorol. Z.* (2006) 259–263, <https://doi.org/10.1127/0941-2948/2006/0130>.
- [48] P. Panagos, The European soil database, *GEO Connex.* 5 (2006) 32–33. <https://esdac.jrc.ec.europa.eu/>.
- [49] FAO, 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/>, 2003.
- [50] Eurostat, Statistical Regions in the European Union and Partner Countries. NUTS and Statistical Regions 2021, 2022 edition, 2022. Eurostat 188, <https://ec.europa.eu/eurostat/web/nuts/background>.
- [51] S. Siebert, J. Burke, J.M. Faures, K. Frenken, J. Hoogeveen, P. Döll, F.T. Portmann, Groundwater use for irrigation – a global inventory, *Hydrol. Earth Syst. Sci.* 14 (2010) 1863–1880, <https://doi.org/10.5194/hess-14-1863-2010>.
- [52] F.T. Portmann, Global Dataset of Monthly Growing Areas of 26 Irrigated Crops: Version 1.0, Univ.-Bibliothek Frankfurt am Main, 2008.
- [53] F.T. Portmann, S. Siebert, P. Döll, 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 (2010), <https://doi.org/10.1029/2008GB003435>.
- [54] H. Ehalt Macedo, B. Lehner, J. Nicell, G. Grill, J. Li, A. Limtong, R. Shakyia, Distribution and characteristics of wastewater treatment plants within the global river network, *Earth Syst. Sci. Data* 14 (2022) 559–577, <https://doi.org/10.5194/essd-14-559-2022>.
- [55] H. Thodsen, H. Tornbjerg, L. Trolborg, J. Windolf, N.B. Ovesen, A. Kjeldgård, A.L. Højbjerg, Udvikling Af Vanddelen Af DK-QNP Til Havbelastningsberegninger. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 2019, 20 s. - Teknisk rapport nr. 145, <http://dce2.au.dk/pub/TR145.pdf>.
- [56] G. Levin, Basemap04: Documentation of the Data and Method for Elaboration of a Land Use and Land Cover Map for Denmark. Technical Report 252. Aarhus, den-mark, Aarhus University, DCE – Danish Centre for Environment and Energy, 2022, p. 77. <http://dce2.au.dk/pub/TR252.pdf>.

- [57] Adhikari, et al., High-resolution 3-D mapping of soil texture in Denmark, English, Soil Sci. Soc. Am. J. 77 (3) (2013) 860–876, <https://doi.org/10.2136/sssaj2012.0275>. ISSN: 0361-5995.
- [58] F. Pianosi, F. Sarrazin, T. Wagener, A Matlab toolbox for global sensitivity analysis, Environ. Model. Software 70 (2015) 80–85, <https://doi.org/10.1016/j.envsoft.2015.04.009>.
- [59] H.V. Gupta, H. Kling, K.K. Yilmaz, G.F. Martinez, Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling, J. Hydrol. 377 (2009) 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>.
- [60] DMI, Climate data. <https://www.dmi.dk/friedata/dokumentation/climate-data>, 2026. (Accessed 18 March 2026).
- [61] C. Plum, G. Martinsen, E. Dybro Thomassen, J. Wied Pedersen, M. Butts, On-demand flood predictions and warnings with high-resolution weather and hydrological models: the case of Vejle, Denmark, EGU General Assembly (2023), <https://doi.org/10.5194/egusphere-egu23-15949>.
- [62] L.S. Matott, OSTRICH: an Optimization Software Tool, Documentation and User's Guide, Version 17.12.19, University at Buffalo Center for Computational Research, 2017, p. 79. <http://www.civil.uwaterloo.ca/envmodelling/Ostrich.html>.
- [63] DHM, Danmarks Højdemodel, 2024. <https://dataforsyningen.dk/data/928>.
- [64] SMHI, Whist. <https://git.smhi.se/whist/whist/-/wikis/WHIST-Wiki/-1.-Introduction-to-WHIST>, 2023.
- [65] C. Brendel, R. Capell, A. Bartosova, Rational gaze: presenting the open-source HYPEtools R package for analysis, visualization, and interpretation of hydrological models and datasets, Environ. Model. Software 178 (2024) 106094, <https://doi.org/10.1016/j.envsoft.2024.106094>.
- [66] L. Rychla, Weekend's torrential rainfalls causing traffic problems, Cph. Post (2015). <https://cphpost.dk/2015-12-28/news/weekends-torrential-rainfalls-causing-traffic-problems>.
- [67] J. Su, J.W. Pedersen, I.M. Ringgaard, M.A.D. Larsen, Incorporating compound flood events in the weather service value chain, Npj Nat. Hazards 2 (2025) 26, <https://doi.org/10.1038/s44304-025-00077-4>.
- [68] J. Liu, J. Koch, S. Stisen, L. Trolborg, R.J.M. Schneider, A national-scale hybrid model for enhanced streamflow estimation – consolidating a physically based hydrological model with long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci. 28 (2024) 2871–2893, <https://doi.org/10.5194/hess-28-2871-2024>.
- [69] J.C.M. Andersson, A.T. Ibrahim, A.L. Soumahoro, V. Fofana, A. Ali, B. Arheimer, How a flood forecasting system saved lives and property in West Africa (no. EGU23-15732), in: Presented at the EGU23, Copernicus Meetings, 2023, <https://doi.org/10.5194/egusphere-egu23-15732>.
- [70] S. Baste, D. Klotz, E. Acuña Espinoza, A. Bardossy, R. Loritz, Unveiling the limits of deep learning models in hydrological extrapolation tasks, Hydrol. Earth Syst. Sci. 29 (2025) 5871–5891, <https://doi.org/10.5194/hess-29-5871-2025>.
- [71] E.A. Espinoza, R. Loritz, F. Kratzert, D. Klotz, M. Gauch, M.Á. Chaves, U. Ehret, Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events, Hydrol. Earth Syst. Sci. 29 (5) (2025) 1277–1294, <https://doi.org/10.5194/hess-29-1277-2025>.
- [72] Y. Song, K. Sawadekar, J.M. Frame, M. Pan, M. Clark, W.J. M. Knoben, A.W. Wood, K.E. Lawson, T. Patel, C. Shen, M.P. Clark, Physics-informed, differentiable hydrologic models for capturing unseen extreme events, Water Resour. Res. 62 (2026) 1–23, <https://doi.org/10.1029/2025WR040414>.
- [73] Y. Hu, H. Li, C. Zhang, T. Wang, W. Chu, R. Li, Investigate the rainfall-runoff relationship and hydrological concepts inside LSTM, Environ. Modell. Software 192 (March) (2025) 106527, <https://doi.org/10.1016/j.envsoft.2025.106527>.
- [74] T. Lees, S. Reece, F. Kratzert, D. Klotz, M. Gauch, J. De Bruijn, R. Kumar Sahu, P. Greve, L. Slater, S.J. Dadson, Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci. 26 (12) (2022) 3079–3101, <https://doi.org/10.5194/hess-26-3079-2022>.
- [75] B. Arheimer, R. Pimentel, K. Isberg, L. Crochemore, J.C.M. Andersson, A. Hasan, L. Pineda, Global catchment modelling using world-wide HYPE (WWH), open data, and stepwise parameter estimation, Hydrol. Earth Syst. Sci. 24 (2020) 535–559, <https://doi.org/10.5194/hess-24-535-2020>.