Development and evaluation of flood forecasting models for forecast-based financing using a novel model suitability matrix

Hagen, Jenny Sjåstad; Cutler, Andrew; Trambauer, Patricia; Weerts, Albrecht; Suarez, Pablo; Solomatine, Dimitri

DOI
10.1016/j.pdisas.2020.100076

Publication date
2020

Document Version
Final published version

Published in
Progress in Disaster Science

Citation (APA)
Hagen, J. S., Cutler, A., Trambauer, P., Weerts, A., Suarez, P., & Solomatine, D. (2020). Development and evaluation of flood forecasting models for forecast-based financing using a novel model suitability matrix. *Progress in Disaster Science*, *6*, Article 100076. https://doi.org/10.1016/j.pdisas.2020.100076

Progress in Disaster Science 6 (2020) 100076

Contents lists available at ScienceDirect

Progress in Disaster Science

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

Development and evaluation of flood forecasting models for forecast-based financing using a novel model suitability matrix

Jenny Sjåstad Hagen [a,b,*], Andrew Cutler [c], Patricia Trambauer [b], Albrecht Weerts [b,d], Pablo Suarez [e,f], Dimitri Solomatine [a,g]

[a] UNESCO-IHE Institute for Water Education, Department for Water Science and Engineering, Westvest 7, 2611 AX Delft, Netherlands
[b] Deltares, Inland Water Systems Division, Boussinesqweg 1, 2629 HV Delft, Netherlands
[c] Boston University, College of Engineering, Department of Electrical Engineering, 8 St. Mary's Street, Boston, MA 02215, USA
[d] Wageningen University and Research, Hydrology and Quantitative Water Management Group, Department of Environmental Sciences, Lumen, Droevendaalsesteeg 3a, 6708 PB Wageningen, Netherlands
[e] Red Cross Red Crescent Climate Centre, Anna van Saksenlaan 50, 2593 HT Den Haag, Netherlands
[f] University College London, Department of Science, Technology, Engineering and Public Policy, 36-37 Fitzroy Square, London W1T 6EY, UK
[g] Delft University of Technology, Water Resources Section, Building 23, Stevinweg 1, 2628 CN Delft, Netherlands

ARTICLE INFO

Article history:
Received 28 November 2019
Received in revised form 6 March 2020
Accepted 7 March 2020
Available online 11 March 2020

Keywords:
Model suitability
Forecast-based financing
Flood forecasting
Neural network
Open data
Delft-FEWS

ABSTRACT

Forecast-based financing is a financial mechanism that facilitates humanitarian actions prior to anticipated floods by triggering release of pre-allocated funds based on exceedance of flood forecast thresholds. This paper presents a novel model suitability matrix that embeds application-specific needs and contingencies at local level on a pilot project of forecast-based financing. The added value of this flexible framework is demonstrated on a set of hydrological and machine learning models. The model suitability matrix facilitates transparency and traceability of subjectivity in model evaluation. This paper advocates a stronger interface between model developers and end users for upscaling of forecast-based financing.

1. Introduction

Since 1900, 55% of globally recorded floods have been classified as river floods [1]. Climate change projections indicate more extreme weather patterns with dry areas getting dryer and wet areas getting wetter; this may accelerate existing flood hazards [2]. Although only low confidence can be given to climate change effects on global flood magnitudes [3], a global assessment of future river flood hazards using eight emission scenarios showed multi-model consistent increases in flood magnitudes across the tropical regions of Africa, South- and East-Asia and Latin-America [4]; these are regions in which intangible flood damage – like loss of lives and spread of waterborne diseases – prevail [5]. Between 2000 and 2009, $<4\%$ of international disaster-related financing was allocated to disaster prevention and preparedness, with the majority of funds allocated to emergency relief [6]. Over time, this imbalance has directed humanitarian aid for disaster risk reduction into two distinct branches, namely emergency relief and long-term disaster risk reduction – leaving a gap in short-term prevention and preparedness.

Operational flood forecasting has become an integral part of flood risk management through wide-spread establishment of early warning systems – at local [7] national [8], continental [9] and global [10] scale. As an extension to flood early warning systems, forecast-based financing is a novel financial mechanism facilitating humanitarian aid prior to anticipated flood events – with practical implications in developing countries, where intangible flood damage prevails. Forecast-based financing consists of three components: i) flood forecast model triggers reflecting local impact levels through forecast thresholds, ii) financial mechanisms which secure and release pre-allocated funds once forecast thresholds are exceeded and iii) a standard operating procedure describing humanitarian actions to be taken by *Red Cross National Societies* and partners once funding is released [11]. By exploiting this "window of opportunity", humanitarian actions – like distribution of water purification

* Corresponding author at: University of Bergen, Geophysical Institute, Allegaten 70, 5020 Bergen, Norway.
E-mail address: jenny.hagen@uib.no. (J.S. Hagen).

http://dx.doi.org/10.1016/j.pdisas.2020.100076

tablets, emergency shelters, canned food and blankets – can prevent loss of life before a flood event turns into a flood disaster. As such, forecast-based financing bridges the existing gap between long-term disaster risk reduction and emergency relief [12].

Although the *International Federation of the Red Cross* has secured funding in the *Disaster Relief Emergency Fund*, forecast-based financing is still limited to pilot projects by 16 *Red Cross National Societies* across Africa, Asia and Latin-America (see https://www.forecast-based-financing.org/our-projects/ for full overview of pilot projects); as goes unsaid, the global effect will increase with upscaling and wide-spread implementation in the Global South. Flood forecasting models can be regarded as the engine of forecast-based financing, but with increasing data availability and open-source code, selecting the most suitable model for forecast-based financing at local level becomes increasingly challenging.

This paper presents the development and subsequent evaluation of flood forecasting models for forecast-based financing. The novelty of this paper comprises a model suitability matrix extending model evaluation beyond the commonly addressed forecast skill. A pilot project of forecast-based financing in Togo, West-Africa, is used as case study, in which the main component of an operational threefold flood forecasting system is improved. First, a process-based distributed hydrological model is set up, calibrated and forced using open and globally available data. Secondly, machine learning models of increasing complexity are trained on local in-situ measurements. Thirdly, a naïve baseline model is defined. Following model construction, a novel model suitability matrix for forecast-based financing is developed and used to evaluate the models. The model suitability matrix considers needs at end-user level through quantitative score assignation on the following criteria: *data*, *software*, *computational efficiency*, *flexibility*, *requirements of technical expertise*, *forecast skill* and *uncertainty*. The aim of this paper is to introduce a holistic and flexible framework for model evaluation targeted to the application of forecast-based financing.

The remaining of this paper is structured as follows: Section 1.1. introduces the case study; Section 2 outlines materials and methods, including development of flood forecasting models (2.1), evaluation of forecast skill (2.2) and evaluation of model suitability (2.3); Section 3 presents the results in terms of forecast skill (3.1) and model suitability (3.2) respectively; Section 4 provides a discussion of the results, with emphasis on the application of the model suitability matrix; and lastly, Section 5 concludes the main findings of the study and gives recommendations for future applications and further development of the model suitability matrix.

1.1. Case study

The Mono River Basin is a transboundary catchment shared between Togo and Benin in West Africa (see Fig. 1.1). The catchment drains an area of 24,100 km^2 between latitude 6°16′N - 9°20′N and longitude 0°42′ E - 2°25′E to the largest river system in Togo: the Mono River. Nangbéto Dam is a medium hydroelectric dam (design capacity: 65 MW) located on the Togolese side of the Mono River Basin (7°25.4′ N, 1°26.1′ E). The dam was constructed in 1987 and is today operated by *Communaute Electrique du Benin*, an electricity company co-owned by Togo and Benin. Downstream of the dam, villages on both Togolese and Beninese sides are exposed to floods on an annual basis during the West African Monsoon (July–October).

The dam reservoir has a retention capacity of 1.72 × 10^6 m^3, but small differences in annual maximum inflow and outflow reveal that water levels are kept high preceding and during the West African Monsoon [13]. Without optimized reservoir release schedules, the reservoir can spill in a matter of days (see Table 1.1) by inflows about the 2-year return period flow or less (see Table 1.2). Table 1.3 shows i) autocorrelation in inflows, ii) autocorrelation in outflows and iii) cross-correlation between inflow and daily average upstream precipitation (from rain gages, not shown), reflecting a slowly responding system in three ways. Firstly, the low cross-correlation between upstream precipitation and inflow to the dam indicates high retention capacity in the soil, so that consecutive, or long-lasting rainfall events are needed to saturate the soil and initiate drainage to the Mono River. Secondly, due to the size of the river system, time is needed for accumulated water to travel downstream. Thirdly, manual

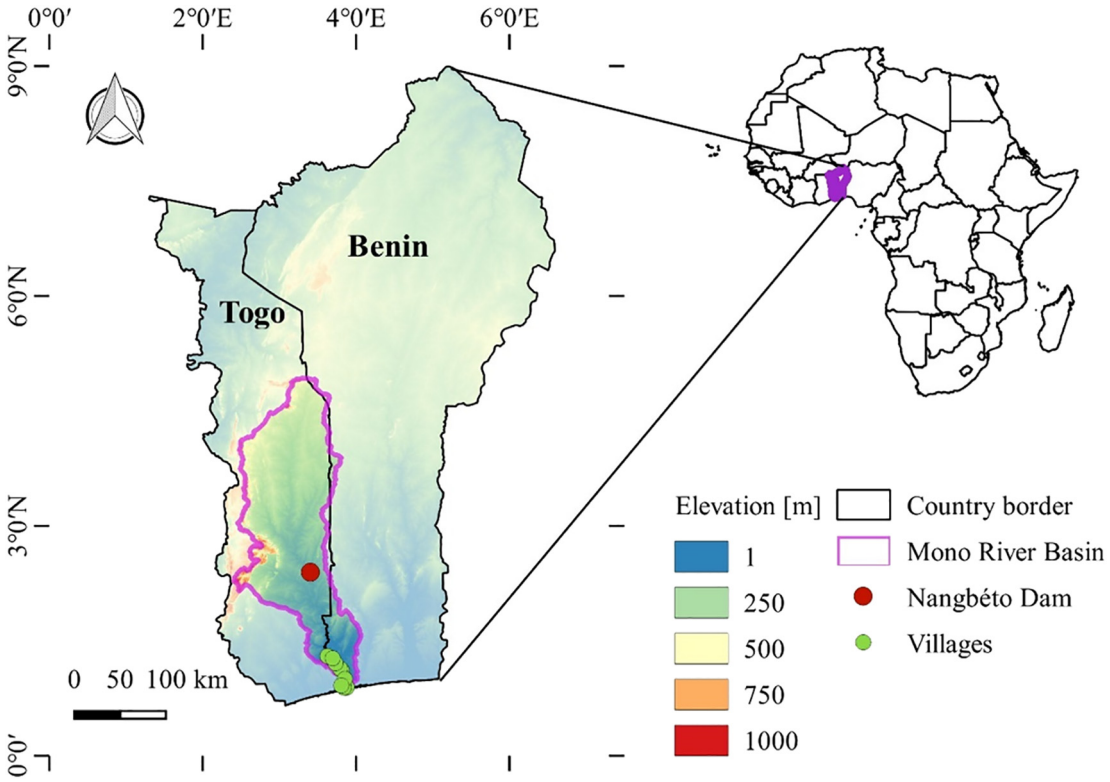

Fig. 1.1. Mono River Basin and key locations (Nangbéto Dam and flood-prone villages).

Table 1.1
Number of days needed to fill the reservoir given constant inflow (Q_{in}) and initial storage as percentage of total volume.

Initial storage	Filling time (days)		
	Q_{in} = 300 m^3/s	Q_{in} = 700 m^3/s	Q_{in} = 1000 m^3/s
10%	60	26	18
20%	53	23	16
30%	46	20	14
40%	40	17	12
50%	33	14	10
60%	26	11	8
70%	20	9	6
80%	13	6	4
90%	7	3	2

release of outflows is dictated by the former to slow processes, with influence from both recent inflows and recent releases (essentially the storage in the reservoir).

Two types of climate are found in the Mono River Basin: sub-equatorial climate with two wet seasons (April–June and September–October) below latitude 8° 30.0′ N, and tropical climate with one wet season (May–September) in the northern parts upstream of Nangbéto Dam. Estimates of mean annual rainfall range between 1060 mm and 1300 mm [14]. The highest rainfall occurs during the West-African Monsoon. Onset of the West African Monsoon is associated with migration of the Intertropical Convergence Zone [16], while variability in rainfall patterns during the West African Monsoon is influenced by the African Easterly Jet (12–15 km altitude) and Tropical Easterly Jet (4–5.5 km attitude) [17], along with local relief and topography (for climatic effects on rainfall in Togo see Ongoma et al. [18]). Following the two wet seasons, a dry season prevails from November to March; <10% of average annual rainfall occurs during these months [19,20]. While declining rainfall rates have been detected between 1960 and 2001 [21], soil-saturation has been identified as the most dominant flood-generating process in the Mono River Basin [22].

Forecast-based financing was operationalized in the Mono River Basin in 2016 by *Togo Red Cross Society*, with support from the *Red Cross Red Crescent Climate Centre*, the *German Red Cross*, *Global Facility for Disaster Reduction and Recovery* and the Togolese Government. The operational threefold flood forecasting system at Nangbéto Dam (the FUNES system) consists of i) an inflow prediction model (FUNES), an outflow prediction model (reservoir model) and iii) a hydraulic model (routing model) (see Fig. 1.2). FUNES [23] is a machine learning model (k nearest neighbor) trained on moving averages of upstream precipitation (7, 14, 21, 56 and 224 days) to predict inflows to Nangbéto Dam with four days lead-time. The predicted inflows are fed to the reservoir model – an exponential function of inflow fitted to historical releases from Nangbéto Dam – and outflows from the reservoir model feed to the routing model. The latter comprises a rudimentary hydraulic model (no measured cross-sections) with subjective probabilities assigned to flood extents in downstream villages. Further reclassification to five risk classes was used to establish triggers for release of pre-allocated funds, secured by the *German Red Cross* with governmental support.

As can be seen in Fig. 1.2, the engine of the operational flood forecasting system is FUNES. The model was transferred to local staff – dam operators at Nangbéto Dam and key persons from the local Red Cross – with the advantage of not requiring high levels of technical expertise. However, since the year of operationalization, FUNES has overestimated inflows, bringing a chain reaction throughout the early warning system that leads to false alarms and subsequent transaction costs (see Fig. 1.3). The dam operators have access to the system and can override flood forecasts that seem highly unlikely, but this would not be needed if the model had higher forecast skill.

Clearly, improvements to FUNES will propagate through the flood forecasting system, reducing false alarms and thereby reducing transaction costs. Therefore, FUNES is subject to improvements through development of a collection of flood forecasting models in the proceeding section. At the same time, it is desirable to keep requirements of technical expertise at a minimum level, so that locally available and affordable levels of expertise are required to implement, operate and maintain the model.

Table 1.3
Autocorrelation in inflows (Q_{in}) and outflows (Q_{out}) and cross-correlation between inflows and daily average cumulative upstream precipitation (Q_{in}-P) for lag days T_{-x}.

	T_{-1}	T_{-2}	T_{-3}	T_{-4}	T_{-5}	T_{-6}	T_{-7}	T_{-8}	T_{-9}	T_{-10}
Q_{in}	0.95	0.92	0.89	0.86	0.84	0.82	0.81	0.79	0.78	0.77
Q_{out}	0.95	0.90	0.87	0.84	0.81	0.78	0.76	0.73	0.71	0.69
Q_{in}-P	0.29	0.33	0.39	0.39	0.38	0.39	0.39	0.35	0.35	0.34

2. Material and methods

2.1. Development of flood forecasting models

The plethora of available software for hydrological prediction is growing, unlocking opportunities for increasingly sophisticated modelling and further complicating the process of model selection. With respect to increasing code and data availability globally, two model types have been subject to significant advancements over the past decades: process-based distributed hydrological models [24] and machine learning models [25]. Therefore, these two model types were selected for development in this study: a flexible process-based distributed hydrological model for which a global parameter set exists and machine learning models of distinct complexities.

2.1.1. Data

Table 2.1 provides an overview of data used in the study. The data is structured into two categories: “globally available/open data” and “local/purchased data”. The first category refers to data that is free of charge and/or obtained from datasets covering the entire globe, derived in such a way that availability is insensitive to geographical location and/or inarguably accessible without costs. Hence, globally available/open data is distinctly different from in-situ measurements owned and protected by local agents and globally distributed data available at a cost. The second category refers to data that is available for specific locations – such as in-situ measurements – and/or at a cost. Local/purchased data was obtained from the *Red Cross Red Crescent Climate Centre* (*RCCC*) and partners associated with implementation of flood forecasting for the pilot project of forecast-based financing in the Mono River Basin.

A comparison of local rainfall measurements and globally available/open data climatology showed that MSWEP reanalysis rainfall

Table 1.2
Flow return periods estimated for dam inflows (Q_{in}), outflows (Q_{out}) and a river gauge station 150 km downstream (Q).

Return period (years)	2	5	10	20	25	50	100	200	Function	Source
Q_{in} (m^3/s)	960	1330	1580	1815	–	2125	2370	–	Fréchet	[13]
	940	1290	1500	–	1740	1910	2060	2210	Log Pearson Type 3	This study
Q_{out} (m^3/s)	530	915	1219	1530	–	1990	2390	–	Weibull	[13]
	455	870	1223	–	1775	2260	2823	3460	Log Pearson Type 3	This study
Q (m^3/s) [annually flooded]	570	800	880	940	–	1000	1040	1780	Gumbel	[14]
	630	850	1030	1200	–	1440	1600	1070	Goodrich	

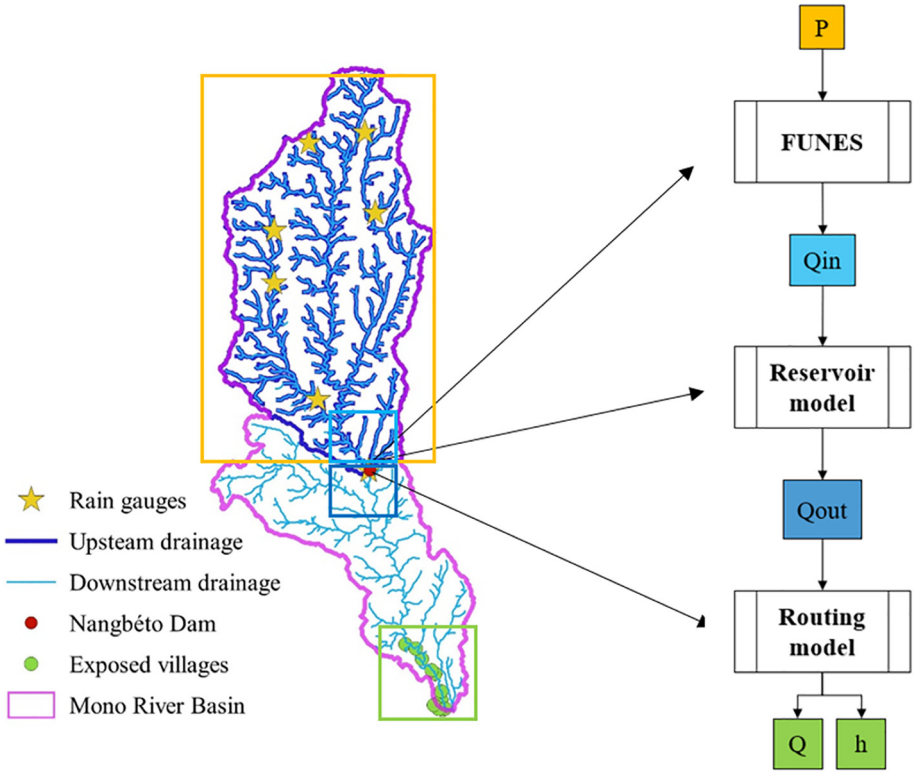

Fig. 1.2. Current threefold flood forecasting system (FUNES scheme) showing models and input/output variables: precipitation (P), inflow to the dam (Q_{in}), outflow from the dam (Q_{out}) and water level (h). The input/output variables as well as the models are located in the catchment with colors and arrows respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

overestimates rainfall over the Mono River Basin (annual average rainfall range 1200–1600 mm), with less bias in the RFE satellite rainfall estimates (annual rainfall range 1100–1300 mm). MSWEP and RFE were initially chosen because these products were found most reliable over West-Africa in previous studies [36]. Table 2.2 provides an overview of statistics of the data used for training and testing of machine learning models.

2.1.2. Hydrological model

The distributed hydrological modelling platform, wflow, is an open-source toolkit of the Deltares Open Streams Project [37]. Wflow currently contains four hydrological models (wflow_sbm [38]; wflow_hbv [39]; wflow_gr4 [40]; and wflow_w3ra [41]. In this study, wflow_sbm (Simple Bucket Model – hereafter referred to as SBM) was used, due to its simplicity and flexibility explained below.

SBM is a modified version of the TOPOG-SBM model, originally developed for steep slopes and thin soil layers (≤ 2 m) by Vertessy and Elsenbeer [42]. As a near calibration-free process-based distributed hydrological model, SBM is designed to maximize information from land cover and soil maps in physically-based parameter estimations. The model is coded in Python-PCRaster [43] and requires i) static input data (digital elevation model, soil and land cover), ii) dynamic input data (precipitation, temperature and potential evapotranspiration) iii) specification of model parameters in PCRaster format. Model parameters are generated with lookup-tables linking soil and land cover to catchment properties. The following processes and fluxes are modelled in response to precipitation: interception, evapotranspiration, infiltration, percolation, horizontal groundwater flow, capillary rise, exfiltration, exchange between groundwater and open water and direct runoff. Lakes, dam reservoirs and irrigation fields can be added to the basic model structure. SBM has been used for hydrological assessments of land cover change [44], benchmarking of global hydrological models in river basin modelling [45] and flood forecasting [31] using both local/purchased and globally available/open data in Africa, Asia, Latin-America, Europe and Australia. SBM was selected for this study as it reflects state-of-the-art process-based distributed hydrological models with kinematic routing of surface water.

Delft-FEWS (Flood Early Warning System) is a data-centric open shell facilitating data handling and forecasting [46]. Delft-FEWS consists of a database, a general adapter, import/export and transformation modules and a user interface. Delft-FEWS has been applied in >40 flood forecasting centers and is currently in operational use in the UK [8], Australia [47] and the USA [48]. Although intended for operational application, the use for research purposes has also been demonstrated [49]. SBM was set up and embedded in Delft-FEWS using the Delft-FEWS Accelerator. To minimize calibration efforts, seamless large-domain parameter estimates [31] developed for SBM were used to generate PCRaster maps of saturated hydraulic conductivity, monthly leaf area index, saturated and residual soil water content, saturated water fraction (lakes), land cover and soil depth. A reservoir was specified at the location of Nangbéto Dam, using estimated reservoir dimensions and a target release (for power production).

The satellite and reanalysis data was merged in Delft-FEWS to create continuous records between 1987 and 2018, in which RFE/ERA5 data was used where available and MSWEP/EartH2Observe given secondary priority. Simulations were run on three-hourly time-steps between 1987 and 2018. Potential evapotranspiration was calculated with de Bruin Equation [50] using merged mean sea level pressure, incoming solar radiation and air temperature. The potential evapotranspiration climatology was calculated from simulations and used for forecasting. Forecasts were generated on six-hourly time-steps from 2016 to 2018 with up to ten days lead-time using ensemble weather forecasts from the ECMWF Ensemble Prediction System (one control forecast and 50 ensemble members). Daily average flows were calculated with a simple averaging procedure in Delft-FEWS.

2.1.3. Machine learning models

Several machine learning models were built, from which the simplest and most complex models were selected. The models were trained using

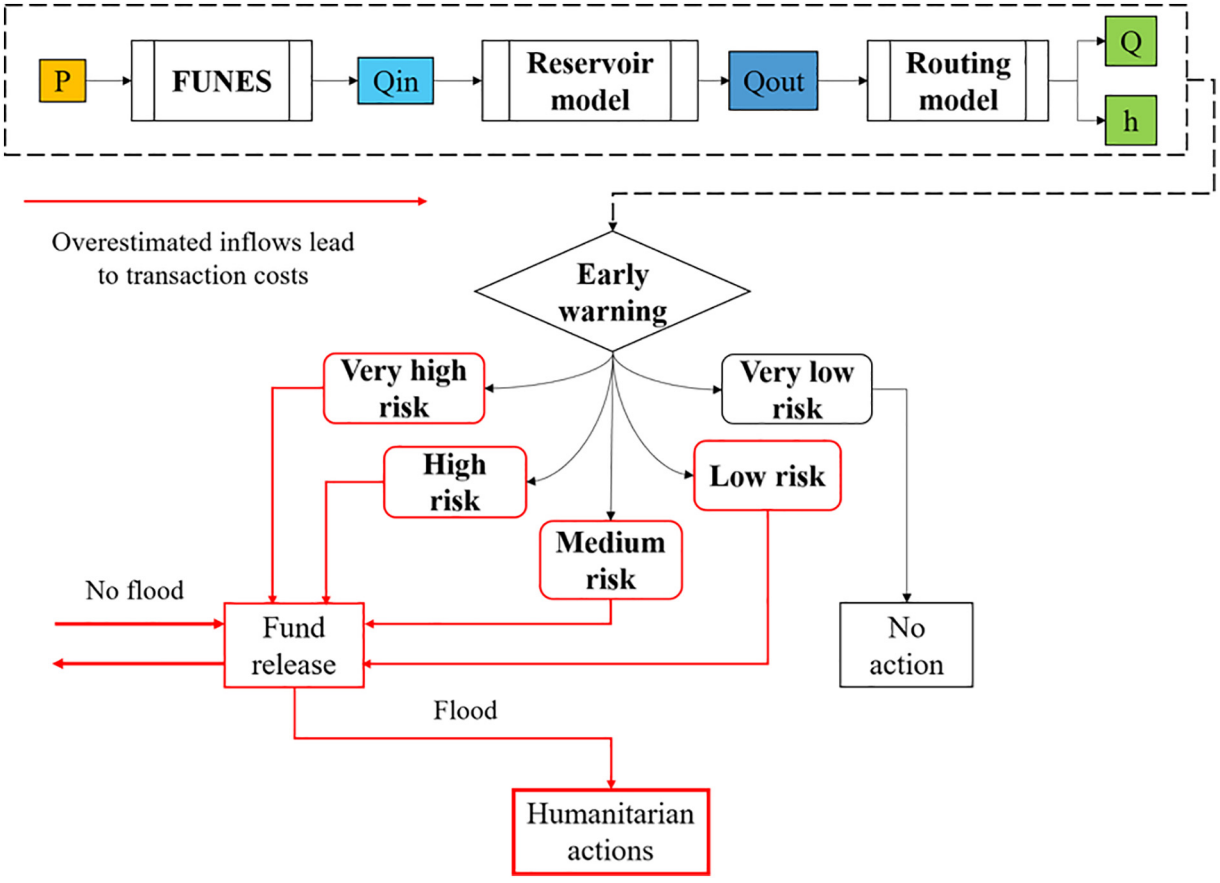

Fig. 1.3. The structure of the FUNES system, currently operational for forecast-based financing in Togo. P = precipitation, Q_{in} = inflow, Q_{out} = outflow, Q = discharge and h = water level.

average-based metrics that emphasize model simplicity and accuracy. The input variable selections were defined based on correlation analyses. Low correlation between rainfall and discharge indicates a slowly responding basin, in which autocorrelation in flows may be exploited for predictive capacity (recall Table 1.3). Using flow data from the preceding 50 days, an autoregressive random forest was trained to predict inflows to Nangbéto Dam. From a supervised learning perspective, this is a reasonable attempt at exploiting the autocorrelation. Following this, several feedforward and feed-backward neural networks were trained using backpropagation with variations of this input variable selection (including the difference in inflows and outflows over consecutive days). Finally, a deep learning model was built, trained on inputs to the hydrological model (precipitation, temperature and potential evapotranspiration) in addition to inflows and outflows. The simplest and most complex machine learning models were used to investigate performance gains from increasing model complexity in terms of architectures and data. The models and experimental setup are described in detail below.

A random forest (RF) is an ensemble of weak classifiers (trees) that average their predictions. Each tree sorts similar samples into groups. At test time, trees can assign labels by averaging nearby samples in the training data. In this case samples are sorted by the 50 preceding days of inflow and outflow and the predicted label is the next ten days of flow (see Table 2.3).

Table 2.1
Overview of data classified according to cost and availability.

	Data	Source	Spatial resolution	Temporal resolution	Period
Globally available/open data	Reanalysis rainfall estimates	MSWEP [26]	28 km	3 h	1987–2015
	Satellite rainfall estimates	RFE [27]	28 km	24 h	2001–2018
	Reanalysis temperature	EartH2Observe [28]; ERA5 [29]	28 km	3 h; 1 h	1987–2015; 2008–2018
	Reanalysis incoming solar radiation	EartH2Observe [28]; ERA5 [29]	28 km	3 h; 1 h	1987–2015; 2008–2018
	Reanalysis mean sea level pressure	EartH2Observe [28]; ERA5 [29]	28 km	24 h; 1 h	1987–2015; 2008–2018
	Ensemble rainfall forecasts	ECMWF-EPS [30]	50 km	6 h	2016–2018
	Global parameter set for SBM	Github [31]	–	–	–
	Land cover	USGS Land Cover Institute [32]	300 m	–	–
	Soil	Harmonized World Soil Database [33]	1 km	–	–
	Streamlines and basin boundaries	HydroBasins [34]	500 m	–	–
	Digital elevation model	SRTM [35]	30 m	–	–
Local/purchased data	Discharge measurements	Nangbéto Dam	Inflow and outflow	24 h (average)	1987–2018
	FUNES forecasts	RCCC	Inflow	24 h (average)	2016–2018
	Reservoir water level	RCCC	Water level estimates	24 h (average)	2016–2017
	Rainfall measurements	RCCC	8 rain gauges	24 h (average)	2012–2016
	Other discharge measurements (incomplete)	RCCC	7 river gauges	24 h (average)	1951–1999 (large gaps)

Table 2.2
Statistical moments of inflows (Q_{in}) and outflows (Q_{out}) used as training and testing data for machine learning models.

	Training (1987–2016)		Testing (2016–2018)	
	Q_{in}	Q_{out}	Q_{in}	Q_{out}
Mean	103 m^3/s	94 m^3/s	131 m^3/s	119 m^3/s
Min	0 m^3/s	0 m^3/s	0 m3/s	0 m^3/s
Max	2133 m^3/s	1429 m^3/s	1391 m^3/s	861 m^3/s
Standard deviation	201 m^3/s	119 m^3/s	235 m^3/s	145 m^3/s
Skewness	9.3	28	4.1	6.6
Kurtosis	2.8	4.4	2.2	2.3

The full existing dataset of daily inflows and outflows at Nangbéto Dam since the year of construction (1987) to 2018 was split into training and verification with statistical justification: The first 90% of the dataset were used for training (1987–2016) and the remaining 10% (2016–2018) were used for testing. To the extent that errors of individual trees are uncorrelated, the ensemble will be more accurate than any one classifier. To this end, bootstrap sampling (training examples drawn with replacement) is used to create classifiers that have seen slightly different sets of data. The model was implemented in the Python library *sklearn* using *ExtraTreesRegressor* [51]. For details on similarity and sorting see Geurts et al. [52], where it is also shown that random forests are a type of k-nearest neighbor model (like FUNES) with weighted voting by neighbors.

Deep learning has the advantage of being able to scale and combine different types of data to make predictions. A convolutional neural network (CNN) is a feature extractor that emphasizes and pools essential information while preserving spatial and temporal components, such as georeferenced location and time. A CNN was built using the 50 preceding days of inflow and outflow, as well as satellite and reanalysis data including: precipitation (from RFE/MSWEP), temperature (from ERA5) and potential evapotranspiration (derived with the de Bruin Equation) (see Table 2.4), with precedence from merging procedure in Delft-FEWS as described in Section 2.1.2. The satellite and reanalysis data were fed to three separable convolutional neural networks [53], and the flow data was fed to a one-dimensional convolutional neural network (as described in LeCun and Bengio [54]). The output of these four networks were concatenated and followed by a fully connected network that predicts inflow to the dam. The architecture is sketched in Fig. 2.1. The model was implemented using the SeparableConv2D model from the Python library *Keras* [55].

While neural networks are the most widely applied machine learning technique in the fields of hydrology and hydraulics [56,57], neural networks are far more complex and sensitive to parameters than random forests. The resulting CNN model had ~24,000 parameters, while 90% of historical data from 1987 gave ~10,000 examples; fitting a model with more parameters than training examples is an ill-posed problem. Stochastic Gradient Descent (SGD) is an iterative method that incrementally updates the parameters using partial derivatives. Upon random weight initialization, the contribution to prediction error per parameter is calculated for a single sample to guide incremental changes in parameter values. For a given set of identical predictions on training data, numerous parameter configurations can produce equal outputs. This equates to the issue of equifinality in conventional hydrological modelling [58]. Equifinality is a common problem when training neural networks and can often be solved by two simple regularization methods:

Table 2.3
Input data to the random forest (subscripts designate preceding days) using flow measurements between 1987 and 2016.

Input						Output				
Q_{in-49}	Q_{out-49}	Q_{in-48}	Q_{out-48}	…	Q_{in}	Q_{out}	Q_{in+1}	Q_{in+2}	…	Q_{in+10}

Table 2.4
Input data to the convolutional neural network. The precipitation (P), potential evapotranspiration (PET) and temperature (T) covered the whole basin.

Input						Output					
Q_{in-49}	…	Q_{in}	Q_{out-49}	…	Q_{out}	Q_{in+1}	Q_{in+2}	Q_{in+3}	…	Q_{in+10}	
P_{-49}		…			P						
PET_{-49}		…			PET						
T_{-49}		…			T						

1. **Weight penalty**: The network is discouraged to use all the ~24,000 parameters by introducing a cost function for the use of each additional parameter. In other words, only parameters that aid learning of multiple examples will be applied; the model is less likely to use parameters to memorize the flow characteristics of a single example [59].
2. **Dropout**: At each step in SGD, a percentage (20–50%) of the weights are temporarily set to zero. The remaining weights (80–50%) of the model must then be able to make flow predictions independent of the weights that have been dropped out. This makes it more difficult for a model to memorize the training set without learning general patterns [60].

In combination with SGD, these regularization methods perform the task of selecting, scaling and combining input variables into higher-order features as the data moves through the neural network. In this study, only weight penalty was applied in combination with SGD, as further improvements were not obtained using dropout.

2.1.4. Baseline model

A model should be as simple as possible – but not simpler. The simplest model used as baseline in this study is a naïve forecast predicting that the measured inflow of today persists k (k = lead-time) days into the future. This model was used as a transparent reference to assess the gain in forecast skill with increasing model complexity.

2.2. Evaluation of forecast skill

Nash Sutcliffe Efficiency (NSE) (Eq. (1)) and root-mean squared error (RMSE) (Eq. (2)) are among the most commonly applied metrics for evaluation of hydrological models. However, since limitations are evident with any single metric, a combination of absolute value error statistics (such as RMSE), normalized goodness-of-fit statistics (such as NSE) and graphical results is recommended [61]. The Kling Gupta Efficiency (KGE) (Eq. (3)) was introduced as a decomposition of the NSE to correlation, bias and variability, but nevertheless suffers from limitations of absolute value error statistics. The Index of Agreement (A_{INDEX}) (Eq. (4)) states the ratio of mean squared error to potential error and partially overcomes the insensitivity of NSE to observed and predicted means and variances. However, poor model fits can obtain high values (>0.65), ultimately precluding calibration with a narrow range [62].

$$NSE = 1 - \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (P_i - \overline{P})^2}, \quad (1)$$

where $P = predicted, O = observed, n = sample\ size$ *and bars denote mean*

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}}, \text{where } P = predicted, O = observed \text{ and } n = sample\ size \quad (2)$$

$$KGE = 1 - \sqrt{(r^2 - 1)^2 + \left(\frac{\sigma_P}{\sigma_O} - 1\right)^2 + \left(\frac{\overline{P}}{\overline{O}} - 1\right)^2}, \text{where } r^2 = coefficient\ of\ determination, \sigma = standard\ deviation \text{ and bars denote mean} \quad (3)$$

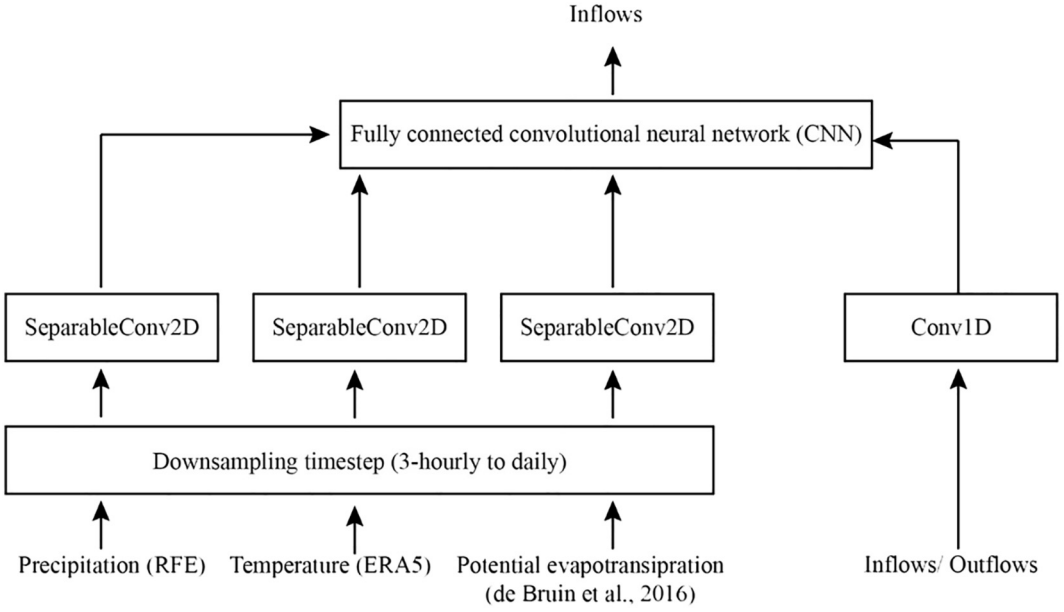

Fig. 2.1. Architecture of the convolutional neural network with precedency of precipitation and rainfall inputs specified in parentheses.

$$A_{INDEX} = 1 - \frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} \left(|P_i - \overline{O}| + |O_i - \overline{O}| \right)^2}, \quad (4)$$

where $P = predicted, O = observed, n = sample\ size$ *and bars denote mean*

While RMSE, NSE, KGE and A_{INDEX} reflect different quantities, they are all average-based metrics. A major limitation of such metrics is that a model may perform well on average and still under- or overestimate flows on a daily basis. In the context of forecast-based financing, where the exceedance of forecast thresholds on daily basis constitutes a main pillar of the system, capturing average performance is insufficient. As argued by Coughlan de Perez et al. [13], the use of hit rate (HR) (Eq. (5)) and false alarm rate (FAR) (Eq. (6)) is advised. However, systematic overestimation can lead to misleadingly high HR, and the number of observed threshold exceedances in the verification period affects the corresponding FAR. It is therefore argued here that a combination of average-based metrics and HR/FAR be used for evaluation of forecast skill; firstly, to constrain predicted flows to the range of observed flows, and secondly, to ensure that the model differentiates flows above and below the forecast threshold. To isolate the forecast skill of the models during the flood season, the metrics were calculated for high flows (West-African Monsoon) and low flows (dry period) separately in the period of verification (2016–2018).

$$HR = \frac{a}{a + c}, \quad (5)$$

where $a = threshold\ exceedance\ forecasted\ and\ observed,$

$c = threshold\ exceedance\ observed\ but\ not\ forecasted$

$$FAR = \frac{b}{b + d}, \quad (6)$$

where $b = threshold\ exceedance\ forcasted\ but\ not\ observed,$

$d = threshold\ exceedance\ neither\ forecasted\ nor\ observed$

Evaluation of probabilistic forecasts was carried out with the Ensemble Verification System (EVS) version 5.6 [63], using the following metrics to capture resolution, reliability and discrimination of forecast probabilities: the Brier skill score (BSS) (Eq. (7)), the mean continuous ranked probability skill score (CRPSS) (Eq. (8)) and the relative operating characteristic score (ROCS) (Eq. (9)). The forecast probabilities were verified against the sample climatology. The advantage of using BSS, ROCS and CRPSS as opposed to the BS, ROC and CRPS is that the ensemble skill (and not just the ensemble spread) is evaluated with reference to the sample climatology. In order to assess the ensemble spread, rank histograms were used. Rank histograms are constructed by counting the fraction of observations that fall in n + 1 ranked ensemble members (bins) and comparing those to a uniform probability across all bins.

$$BSS = 1 - \frac{BS}{BS_{ref}}, where\ BS = \frac{1}{n} \sum_{i=1}^{n} (P_i - O_i)^2, \quad (7)$$

$P = forecasted\ probability, O = observed,$

$n = sample\ size$ *and subscript ref denotes the sample climatology*

$$CRPSS = 1 - \frac{CRPS}{CRPS_{ref}}, where\ CRPS = \int_{-\infty}^{\infty} \sum_{i=1}^{n} \frac{(P_i - O_i)^2}{n} dP, \quad (8)$$

$P = forecasted\ probability, O = observed,$

$n = sample\ size$ *and subscript ref denotes the sample climatology*

$$ROCS = 1 - \frac{ROC}{ROC_{ref}}, \quad (9)$$

where $ROC = 2(A - 0.5),$

where $A = area\ under\ curve\ obtained\ by\ plotting\ HR\ against\ FAR$

Yet, models obtaining high forecast skill can vary widely on other aspects impacting operability, such as complexity, flexibility and data – and code availability. In the context of forecast-based financing, it can be argued that the criteria for model evaluation be extended beyond forecast skill to consider local contingencies and needs at end-user level. This can be obtained by applying a novel approach to model evaluation using the model suitability matrix presented in the proceeding section.

2.3. Evaluation of model suitability

Subjective decisions, or opinions embedded in mathematics, are intrinsic to both model development and model evaluation. The model at hand is optimized to perform well on selected metrics that essentially reflect judgements made during model development. This subjectivity is however often explicitly or implicitly undermined; the authors therefore

argue here that a transparent approach for tracing quantified opinions is needed. When quantified opinions can be traced, strengths and weaknesses can be identified according to end-user needs; the model suitability matrix is an attempt at this.

In defining the model suitability matrix for forecast-based financing, the following generic steps were taken:

1. Define criteria of interest
2. Select metrics and suitability thresholds for quantitative score assignation
3. Embed suitability thresholds in decision tree for transparency
4. Select forecast lead time and use decision tree to consistently assign scores
5. Normalize scores in suitability matrix and display in radar charts

The above-mentioned steps are generic in the sense that various stakeholder constellations can follow the same procedure to adapt the framework for model evaluation on a case-by-case basis. As such, the framework is flexible, transparent and consistent. The criteria and thresholds presented below reflect expert judgement by the authors and are meant to illustrate the setup of the framework rather than provide solid numbers for other case studies and future applications. Seven criteria were defined in collaboration with representatives from the *Red Cross Red Crescent Climate Centre* affiliated with the pilot project of forecast-based financing in the Mono River Basin (see Table 2.5).

Data availability and costs are the largest constraints for implementing flood forecasting systems in developing countries with data-sparse catchments – where forecast-based financing is most needed. Likewise, financial constraints can restrict the use of commercial software. Therefore, the use of freely available data (ID1) and open-source code (ID2) is promoted. Computational efficiency (ID3) relates to resources needed to run the model and connects to lead time in the sense that the time needed to generate forecasts affects the lead-time of the forecast in real-time. Therefore, low computational efficiency is rewarded.

As catchment characteristics change and data availability increases, models should be able to cope with and benefit from such changes (ID4). Models that allow for data assimilation and incorporation of, for instance, land cover changes or dam reservoirs upstream or downstream are therefore considered more flexible. Moreover, models should require an obtainable level of technical expertise among local staff for operation and maintenance (ID5). Forecast-based financing is a mechanism that is handed over from the training agency (Red Cross Red Crescent Climate Centre and partners) to local staff (local authorities or other emergency management first responders and local representatives of the National Red Cross Societies). This transfer is usually through in-person interaction, for instance through a workshop. Given limited time and resources, a workshop of one week was considered a reasonable estimate for a fast transfer to illustrate the use of this criterion. However, while a training time for local staff of one week here is used as a proxy for an efficient transfer, it should be noted that this will vary from case to case, depending on available resources.

Table 2.5
Selected criteria with ID reference and description.

ID	Criteria	Description
1	Data	The degree to which data used in model setup is available and free of charge regardless of geographical location.
2	Software	The degree to which open-source code comprises the model structure.
3	Computational efficiency	The time required to generate forecasts relative to the forecast lead-time.
4	Flexibility	The degree to which the model can adapt to catchment changes and incorporate observations through data assimilation.
5	Requirements of technical expertise	The time needed for untrained local staff to acquire technical skills and knowledge needed to operate and maintain the model independently from model developers.
6	Forecast skill	Accuracy expressed in terms of hit rates, false alarm rates and average-based metrics like NSE, RMSE, KGE and A_{INDEX}.
7	Uncertainty	The degree to which forecast uncertainty is displayed in model outputs.

As argued above, the combined use of average-based metrics and HR/FAR is more appropriate for forecast-based financing, as release of funding is triggered once forecast thresholds are exceeded (ID6). Lastly, all models exhibit uncertainty from input, structure and parameters, but not all display this uncertainty in model outputs. By using models that display uncertainty (ID7), probabilities – rather than a single deterministic value – can form basis for forecast thresholds defined to trigger release of funding for forecast-based financing. This aligns forecast-based financing with the recent shift from deterministic to probabilistic flood forecasting [64]. In this paper, the BSS, CRPSS and ROCS described above were used to assess the performance of the probabilistic forecast against the sample climatology. For BSS > 0, CRPSS> and ROCS > 0, the sample climatology is considered outperformed by the probabilistic forecast, so that the uncertainty displayed by the ensemble provides an added value and hence one point is obtained on ID7 (see Fig. 2.2). The three scores, BSS, ROCS, and CRPSS, were used complimentarily, but other metrics may also be utilized complimentarily or separately to assess the performance of the forecast ensemble against the sample climatology.

The decision tree guiding score assignation on the seven criteria specified above is presented in Fig. 2.2. The questions were formed and structured in such a way that scores were assigned according to the relative importance considered by the authors. After score assignation, the model suitability matrix contains scores on each criterion (see Table 2.6 for illustrative setup).

For a visual display of the model suitability, the scores Z per criterion i were linearly normalized between 0 and 1 using minimum/maximum obtainable score Z_{min}/Z_{max} (see Eq. (10)) and thereafter displayed in radar charts.

$$Z_i = \frac{Z_i - Z_{i,\,min}}{Z_{i,\,max} - Z_{i,\,min}} \tag{10}$$

This visualization can be particularly useful when stakeholders with non-technical backgrounds engage in the model selection procedure. It should be stressed that the criteria and thresholds defined above should reflect local contingencies; this can only be obtained through a stakeholder approach. In the proceeding section, the results at four days lead-time are presented. While several models predicted flows with up to ten days lead-time, four days were used for model evaluation in order to compare the results with FUNES.

3. Results

3.1. Forecast skill

Fig. 3.1 shows inflow forecasts during the West African Monsoon (2016/2017) with forecast hits and misses highlighted for the first observed exceedance of the forecast threshold (300 m^3/s). As can be seen, only FUNES obtained a forecast hit at first observed exceedance of the forecast threshold in 2016. However, FUNES consistently overestimated inflows throughout the wet season, causing daily false alarms, and is clearly not constrained within the range of observations.

Given the structure of the baseline model (BAS), the forecast threshold must be observed before it can be predicted; consequently, the first exceedance of the forecast threshold is always missed. The machine learning models (RF and CNN) predicted flows closer to the observations, but missed the first observed exceedance of the forecast threshold in 2016 due to lags and slight underestimation. Despite some over- and underestimation of a smaller magnitude than that of FUNES, the hydrological model (SBM) seems to capture dynamics fairly well relying only on globally available and open data. RF and SBM obtained forecast hits in 2017. The flood peak was larger and occurred earlier in 2017, and

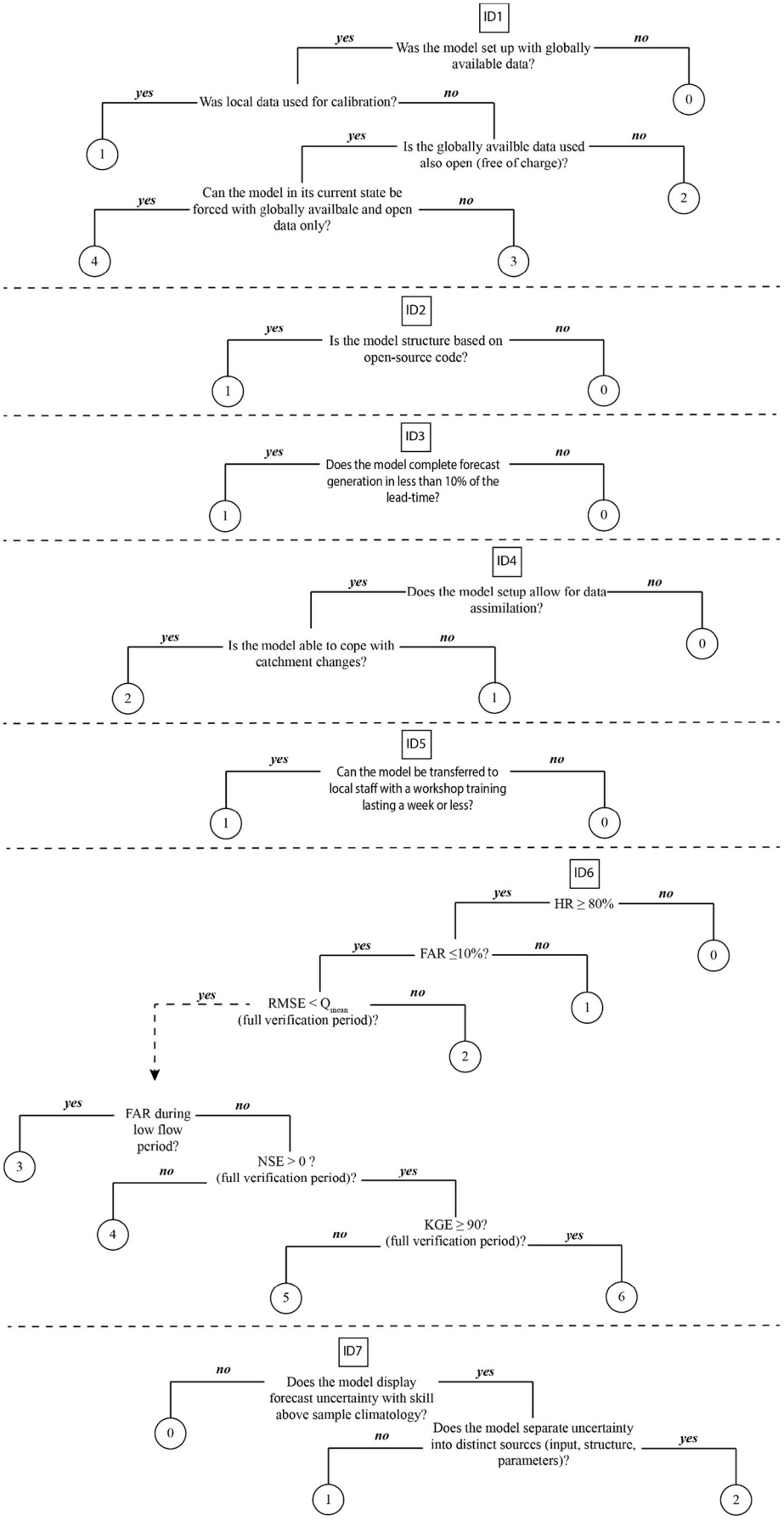

Fig. 2.2. Decision tree for score assignation using number IDs from Table 2.5.

Table 2.6
Setup of model suitability matrix.

Model	Criteria						
	ID1	ID2	ID3	ID4	ID5	ID6	ID7
$Model_1$							
$Model_2$							
…	…	…	…	…	…	…	…
$Model_n$							

while all models except for FUNES obtained forecast hits, the peak was underestimated by all. Contrary to the large overestimation in 2016, FUNES underestimated inflows in 2017, missing the first observed threshold exceedance by several weeks.

The importance of forecast hits at the first observed exceedance of the forecast threshold connects to the fact that effects of humanitarian actions following release of funds often extend beyond the forecast lead-time. While the lead-time constrains the range of actions to be carried out once funding is released, those actions can reduce the existing flood risk in near future; as an example, if water purification tablets and emergency shelters are distributed following a forecast hit that persists into the future, the same actions will not be needed and the risk is reduced until the point in time where all tablets have been used and the shelters been filled. A forecast miss, on the other hand, would allow more damage to occur before actions are taken – essentially turning forecast-based financing into emergency relief.

Scatter plots separated into dry/wet seasons and full verification period, are shown in Fig. 3.2. As can be seen, none of the models performed well during the dry season, and noticeable differences in skill are observed between the West African Monsoon in 2016 and 2017 for all models. Over the full verification period, the machine learning models obtained the highest score, while the naïve forecast outperformed both FUNES and SBM.

Table 3.1 summarizes the hit rates and false alarm rates for the full verification period. The simplest machine learning model (RF) obtained the highest hit rate. The lowest false alarm rate was obtained by the hydrological model (SBM) and the more complex machine learning model (CNN). The baseline model (BAS) obtained the lowest hit rate, but a false alarm rate in the range of RF, CNN and SBM. This is however a reflection of high lag autocorrelation; with large and slowly responding river systems like the Mono River, the advantages of non-autoregressive models like FUNES, CNN or SBM are seen when the forecast lead-time exceeds the period of lag autocorrelation. However, for fair comparison between the models constructed in study, the lead-time was restricted

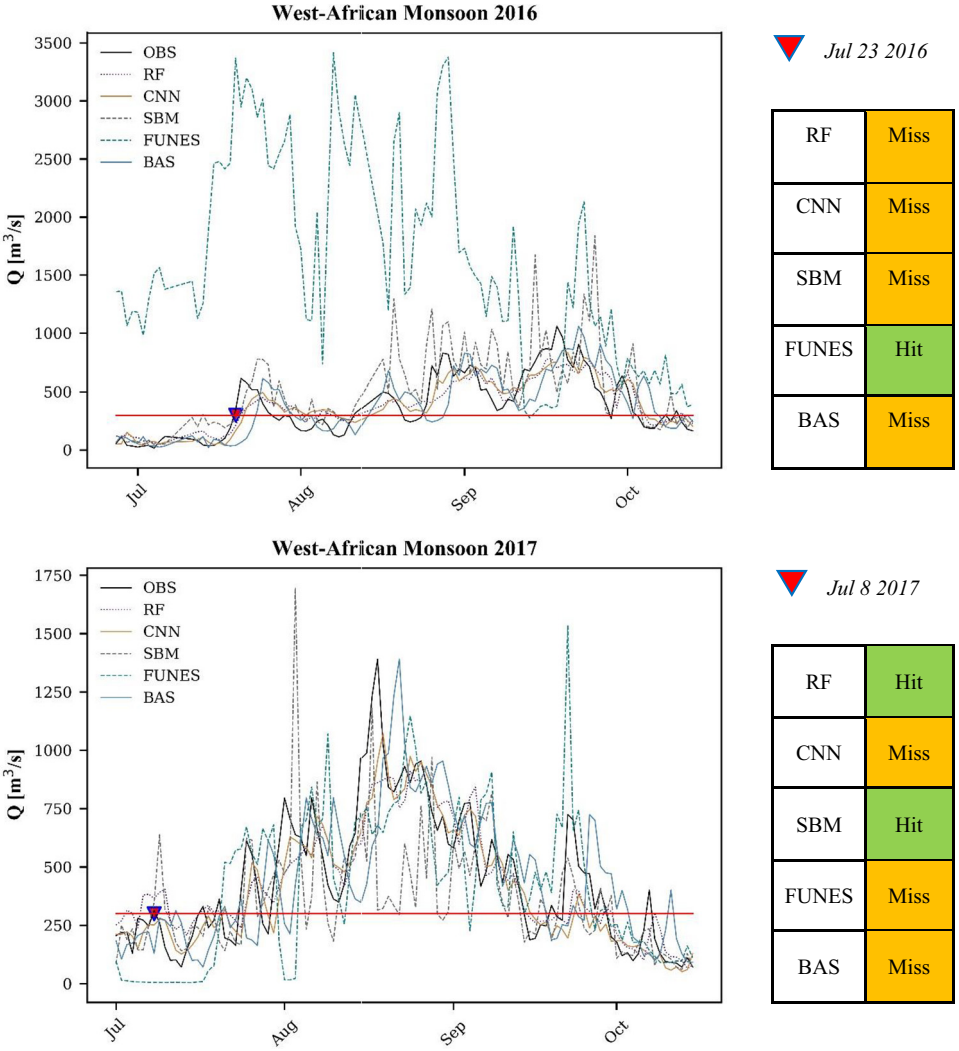

Fig. 3.1. Inflow prediction for the West-African Monsoon in 2016 (top) and 2017 (bottom) by random forest (RF), convolutional neural network (CNN), the hydrological model (SBM), the currently operational model (FUNES) and the baseline model (BAS) as compared to observations (OBS). The red triangle marks the first observed exceedance of the forecast threshold (300 m^3/s) for each flood season and the corresponding forecast hits and misses by the respective models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

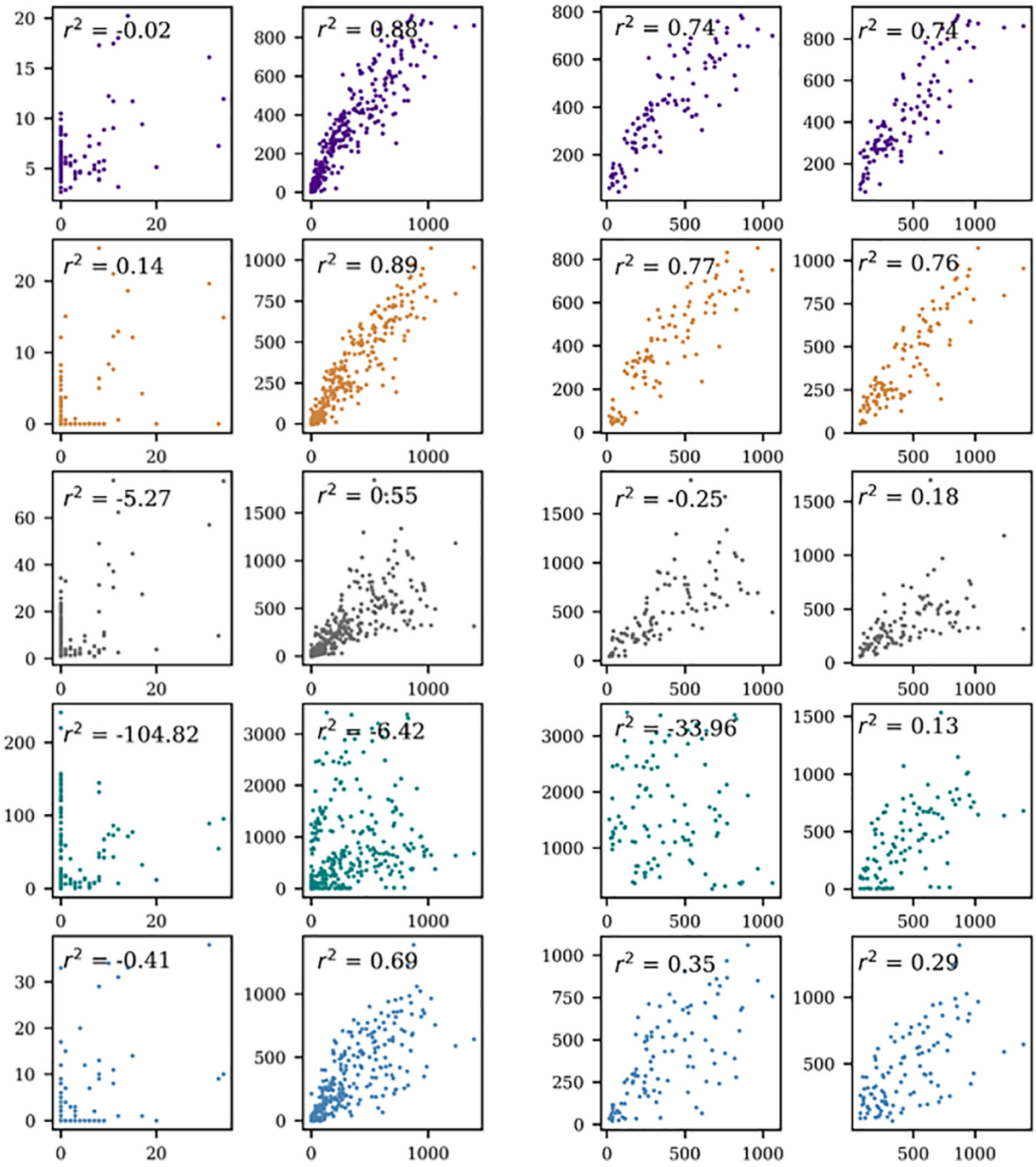

Fig. 3.2. Scatter plots (from left): dry period, full verification period, West-African Monsoon 2016 and West-African Monsoon 2017. The coefficient of determination (r^2) is shown separately for each period.

to four days by the currently operational model FUNES. For the purpose of demonstrating the application of the model suitability matrix, the choice of using four days lead-time does not impinge drawbacks – but the relative forecast skill of the baseline and machine learning models is biased accordingly at lower lead-times. Interestingly, the hydrological model relying solely on globally available and open data obtains a hit rate and false alarm rate of the same order as the machine learning models. However, a clear difference in model performance is seen in terms of absolute value error and goodness-of-fit statistics, where the machine learning models outperform the other models. This is shown in Table 3.2.

Table 3.1
Hit rate (HR) and false alarm rate (FAR) for full verification period.

Model	HR (%)	FAR (%)
RF	91	8
CNN	84	6
SBM	83	6
FUNES	90	23
BAS	78	7

Since the flood forecast models are intended to predict high flows, one can argue that forecast skill during low flows can be disregarded. However, if the forecast threshold is exceeded during the low flow period, this results in false alarms. While neither RF, CNN, SBM nor BAS exceeded the forecast threshold during the dry period of 2017, FUNES exceeded the forecast

Table 3.2
Root-mean squared error (RMSE), Nash-Sutcliffe Efficiency (NSE), Kling-Gupta Efficiency (KGE) and Index of Agreement (A_{INDEX}) for full verification period.

Model	RMSE (m^3/s)[a]	NS (−)	KGE (−)	A_{INDEX} (−)
RF	91	0.88	0.89	0.86
CNN	86	0.89	0.90	0.88
SBM	176	0.55	0.77	0.54
FUNES	718	−6.42	−1.46	−0.26
BAS	146	0.69	0.85	0.66

[a] Q_{mean} = 131 m^3/s (2016–2018).

threshold five times. Table 3.3 summarizes the average-based metrics calculated for the dry period. All models were outperformed by the climatology (observed mean) on one or more metrics.

In terms of resolution, reliability and discrimination, probabilistic SBM forecasts outperformed the sample climatology using the forecast threshold of 300 m^3/s (see Fig. 3.3).

To separate structural errors in the hydrological model from errors in model inputs, the probabilistic forecasts were verified against both observed and simulated flows. As can be seen, the contribution from structural errors is evident at lower lead-times, and the seemingly improving forecast skill with lead-time (up to four days) is a reflection of compensating errors. The convex rank histogram in Fig. 3.4 shows that the ensemble lacks spread as compared to uniform probability, indicating contribution from errors in data inputs (ensemble weather forecast) as well. The ensemble mean did not improve the deterministic SBM forecasts.

3.2. Model suitability

The model suitability matrix is presented in Table 3.4, and the corresponding linearly normalized radar charts are shown in Fig. 3.5. FUNES differs from BAS only in terms of forecast skill (ID6), while RF is a further improvement to FUNES on this criterion. SBM clearly stands out on criteria like data (ID1), software (ID2) flexibility (ID4) and uncertainty (ID7), but due to complexity requires more technical expertise (ID5) and did not obtain the forecast skill of the machine learning models.

Increasing complexity of the machine learning models adds to the forecast skill at the expense of requirements of technical expertise. Hence, CNN obtains the highest forecast skill as evaluated with the model suitability matrix. In terms of model suitability, three distinct groups can be identified. BAS, FUNES and RF are models that are both easy to implement and easily transferred to local staff. However, these models have no to medium high forecast skill. CNN is a complex model with high forecast skill that lacks display of forecast uncertainty and means for data assimilation. SBM is complementary to CNN in the sense that it facilitates probabilistic forecasting with ensemble weather forecasts, allows for data assimilation and incorporation of catchment changes, and can be set up, calibrated and forced with globally available and open data only. BAS is both cheaper and easier to implement than FUNES and RF, but with the hit rate of 78% it failed to obtain forecast skill as defined in the model suitability matrix. Consequently, a combination of CNN and SBM could be used to cover the complementary qualities in a high-tech scenario – or FUNES could be substituted with RF in a low-tech scenario. This is discussed in the proceeding section.

Table 3.3
Root-mean squared error (RMSE), Nash-Sutcliffe Efficiency (NSE), Kling-Gupta Efficiency (KGE) and Index of Agreement (A_{INDEX}) for the dry period.

Model	RMSE (m^3/s)[a]	NS (−)	KGE (−)	A_{INDEX} (−)
RF	6.3	−0.02	−0.27	0.08
CNN	5.8	0.14	0.35	−0.15
SBM	15.6	−5.27	−2.19	0.11
FUNES	63.9	−104	−13.9	−0.15
BAS	7.4	−0.41	0.37	−0.43

[a] Q_{mean} = 3.1 m^3/s (dry period 2017).

4. Discussion: forecast skill vs. model suitability

The discussion on model evaluation in hydrology is not new; rather, it has been going on for decades. In a more general context of science, specialization within sub-disciplines of sub-disciplines have concentrated, narrowed and dogmatized the framework of model evaluation, excluding wider connections to practical applications. By directly involving such connections in the model evaluation, models can be better targeted to the intended practical applications, circumventing operational pitfalls discovered in retrospect and maximizing the utility of the model at hand.

As was demonstrated above, a model with high forecast skill in terms of root-mean squared error or Nash-Sutcliffe Efficiency – both commonly used for evaluation of hydrological models (Dawson et al., 2007) – may still be unsuitable for forecast-based financing, where forecast threshold exceedance is the trigger for early action. Likewise, a model with high hit rate – the metric advocated for forecast-based financing [13] – may still fail to capture general system dynamics and constrain flows within the observed range. The baseline model is an example of the former and FUNES an example of the latter.

The poor performance of FUNES – particularly in 2016 – can be explained by three factors: the sparse data used for training the model before operationalization, the greediness of the heuristic behind the k nearest neighbor algorithm, and the use of inconsistent data sources for training/testing and operationalization respectively. First of all, sparse data limits the ability of the model to generalize. This is not algorithm-specific, but rather a general remark on the use of machine learning; without sufficient data, in terms of both quality and quantity, there is a practical limitation to generalization. Secondly, biases will arise when the k nearest neighbors are not representative – this is a joint consequence of the model structure of FUNES and the limited data, whereby inputs are mapped to outputs by averaging the outputs associated with the k closest inputs. Thirdly, several sources of spatial precipitation were used for training/testing and operationalization. For training, moving averages of measured precipitation upstream of Nangbéto Dam were used. In operational mode, forecasted precipitation is needed. Gridded forecasted precipitation was obtained from several sources, so that biases in inputs may differ, and thereby distorting the signal with more noise.

The main challenge when predicting systems with high autocorrelation is to surpass the naïve forecast. In terms of absolute value error and goodness-of-fit statistics, both machine learning models consistently surpassed the baseline model at four days lead-time. In terms of hit rate and false alarm rate, the hydrological model (SBM) surpassed the baseline model with similar margins as the complex machine learning model (CNN), while the simplest machine learning model (RF) obtained the highest hit rate among all.

While RF, CNN and SBM improved the forecast skill of FUNES to various degrees, they differ significantly in terms of complexity and requirements of technical expertise. The structure of SBM allows for state-updating through data assimilation, and given the role of soil saturation in flood generation in the Mono River Basin [22], it is likely that data assimilation of soil moisture would further improve the forecast skill of SBM. Although CNN obtained the highest forecast skill as defined in the model suitability matrix, it is – like SBM – far more complex than RF. This points in the direction of two distinct possibilities, where an order of precedency is established among the criteria.

If requirements of technical expertise is a less important criterion in the Mono River Basin, the complementary qualities of SBM and CNN argue in favor of a possible hybridization, where for instance a CNN is embedded for error-correction of SBM. This is possible in operational mode with Delft-FEWS. On the other hand, if requirements of technical expertise is a fundamental constraint, it is clear from the model suitability radar charts that RF is a direct improvement to the existing operational model, FUNES, in terms of forecast skill, ceteris paribus. Furthermore, since rainfall rates have declined in the Mono River Basin since 1960 [21] and the autocorrelation in flows is consistently high, it may be more reasonable to

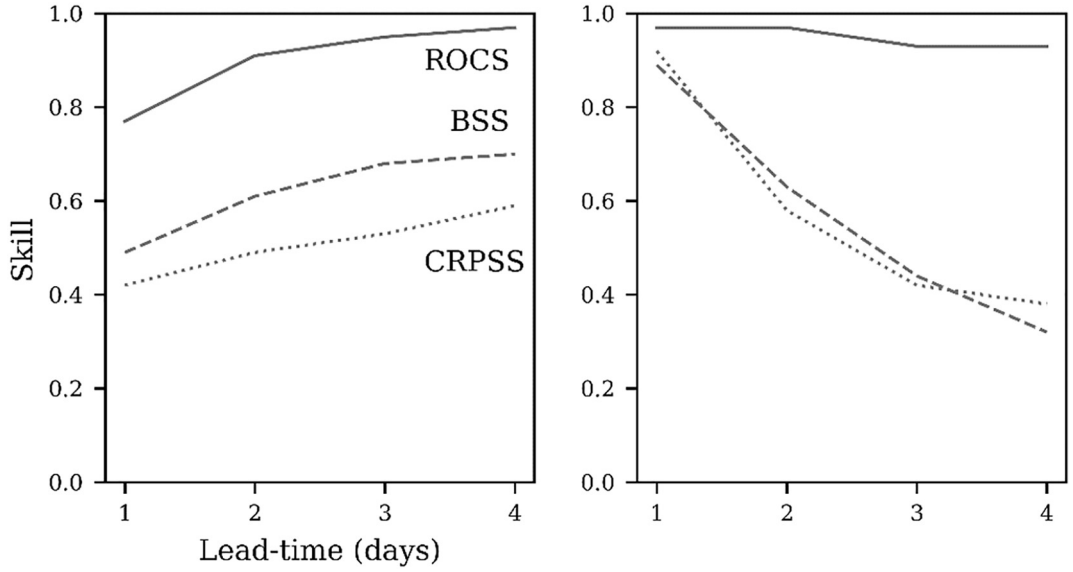

Fig. 3.3. Skill scores of SBM forecast probabilities. Left: Relative operating characteristic score (ROCS), Brier skill score (BSS) and mean continuous ranked probability skill score (CRPSS) compared against the sample climatology. Right: ROCS, BSS and CRPSS compared against simulated inflows.

use flow measurements directly as input data in an autoregressive machine learning model and avoid using non-stationary rainfall records. This can be argued despite the apparent advantages of using forecastable input variables, like precipitation, to obtain forecast skill at lead-times beyond the period of high autocorrelation (approximately 50 days).

While one might argue that process-based distributed hydrological models like SBM could be replaced by a lumped structure for point predictions of inflow at only one location, the advantage of its complexity is that catchment changes can be incorporated into the model structure using the seamless large-domain parameter set [31] for calibration. A second dam downstream of Nangbéto Dam and upstream of the flood-prone villages has been planned for years, and in the case that the dam construction is initiated, inflow forecasts to Nangbéto Dam will lose value for forecast-based financing. SBM is the only model described in this paper that can deal with such changes, for which a second reservoir would be implemented downstream in the distributed grid. Furthermore, land cover changes like urbanization, cultivation or deforestation can be detected with satellite images and readily used to update parameters in the catchment with lookup-tables.

As more data is recorded and made available for model recalibration, FUNES may also improve with recalibration; the effect of this was seen in the forecasts for 2017. However, improvements also depend on the quality of the rain gauge data, the bias in the gridded forecasted precipitation and the stationarity of relevant processes and characteristics in the basin; for instance, in the case of urban development or deforestation upstream, more data from existing rain gauges do not necessarily lead to model improvements in combination with old data, as signals in the new data may add noise to the old data.

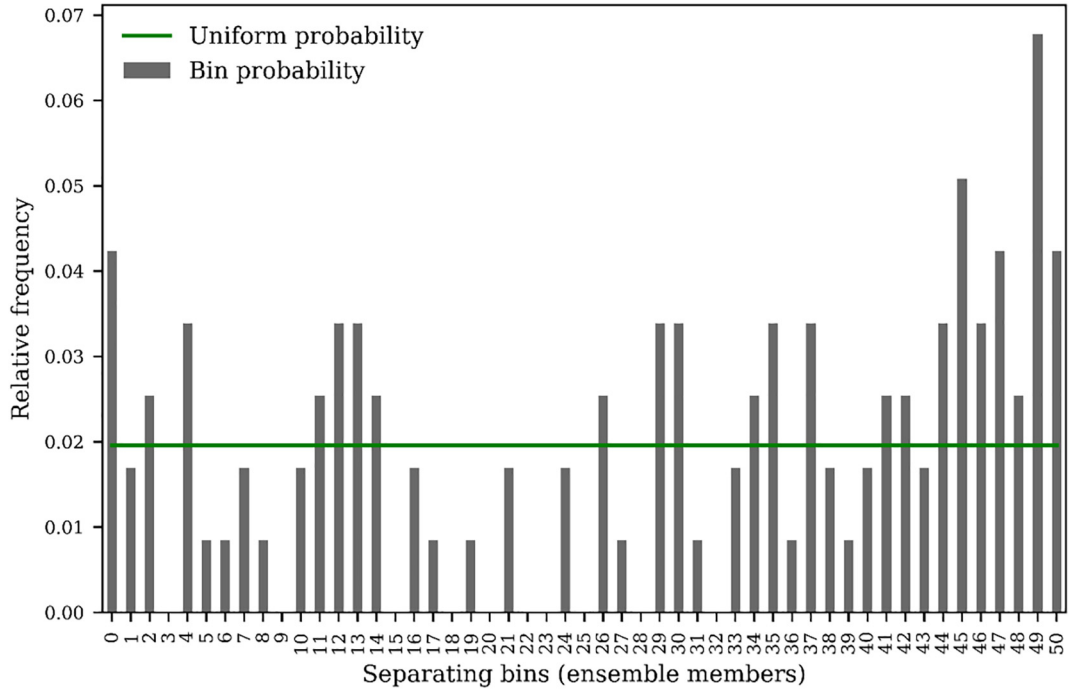

Fig. 3.4. Rank histogram of SBM forecast probabilities at four days lead-time.

Table 3.4
Model suitability matrix showing assigned and minimum/maximum obtainable scores on each criteria.

Model	Criteria						
	ID1	ID2	ID3	ID4	ID5	ID6	ID7
RF	0	1	1	0	1	5	0
CNN	1	1	1	0	0	6	0
SBM	4	1	1	2	0	2	1
FUNES	0	1	1	0	1	1	0
BAS	0	1	1	0	1	0	0
Min.	0	0	0	0	0	0	0
Max.	4	1	1	2	1	6	2

In terms of data and software, all models were set up with open-source code, but only the more complex models, SBM and CNN, were setup and forced with globally available and open data. The use of such data is particularly valuable in developing countries, where continuous and quality-controlled local in-situ measurements often are lacking or difficult to obtain. Although one might argue that the baseline model is the simplest and cheapest implementation, the fact that local in-situ measurements are used makes it less transferable; obtaining such data owned by local agents can be a time-consuming and expensive process, especially if models are developed off -site, and it further requires high autocorrelation in measurements to give the baseline model predictive skill. Lack of data downstream of the dam is the very reason why inflow forecasts to Nangbéto Dam were decided used for forecast-based financing before the pilot project in Togo was operationalized in 2016; however, if a river gauge is set up closer to the flood-prone villages, less data is needed to verify the hydrological model downstream as compared to the machine learning models because the latter model type requires data for both testing and training.

During the dry period, none of the models performed well, and only CNN outperformed the observed mean. It should however be noted that neither the machine learning models nor the hydrological models produced false alarms in the dry period, as opposed to FUNES. The baseline model did not outperform the observed mean. However, as stated before, poor performance during low flows can be neglected if no false alarms are produced. The implication of using a model suitability matrix for model evaluation, as opposed to only looking at forecast skill expressed through a collection of metrics, is that aspects relating directly to model operability can be addressed to facilitate upscaling of forecast-based financing. Concrete recommendations regarding this are given in the conclusion.

5. Conclusion

Using a pilot project of forecast-based financing as case study, this paper presented the development and evaluation of five flood forecasting models: i) a process-based distributed hydrological model (SBM) using globally available data, ii) a simple machine learning model trained on local in-situ measurements (RF), iii) a more complex machine learning model (CNN) additionally trained on satellite precipitation estimates, reanalysis temperature and derived potential evapotranspiration, iv) an operational machine learning model (FUNES) and v) a naïve baseline model. A novel model suitability matrix was introduced, broadening the model evaluation from *forecast skill* to include quantitative score assignation on *data*, *software*, *computational efficiency*, *flexibility*, *requirements of technical expertise* and *uncertainty* with the use of a decision tree. The approach provides a holistic and flexible framework for model evaluation, in which subjective judgements are made transparent and traceable. This contrasts the current practice, where subjective judgements – implicitly or explicitly – are embedded but understated in the process of model development and evaluation. In the context of forecast-based financing, the model suitability matrix allows model developers to embed needs at end-user level and thereby better target the model to its practical application.

For future applications and further development of the model suitability matrix for forecast-based financing, the following recommendations are given:

1. Stakeholders should be engaged in on-going and planned pilot projects of forecast-based financing to develop case-specific model suitability matrices.

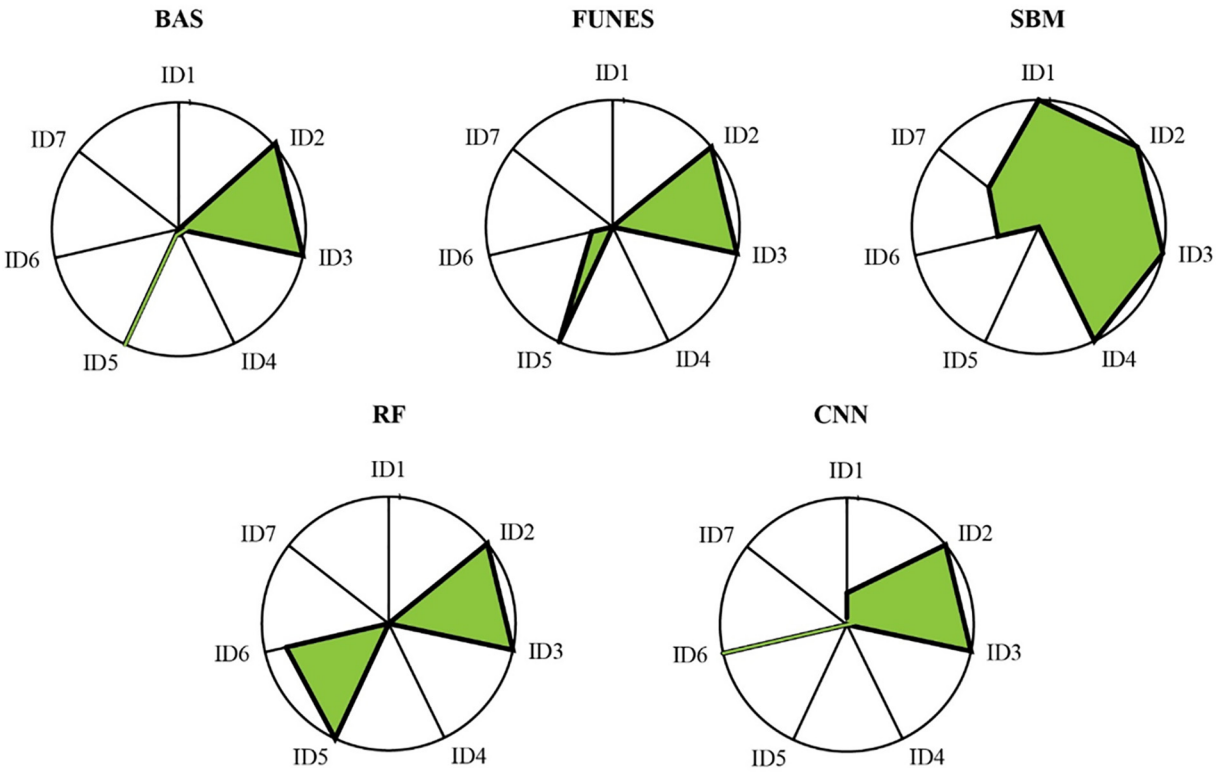

Fig. 3.5. Radar charts displaying model suitability for forecast-based financing in the Mono River Basin.

2. The various model suitability matrices, including criteria and suitability thresholds, should be stored in a database so that future projects can benefit from existing tools based on similarity in terms of local needs, data availability and catchment characteristics; this will support upscaling.
3. The scientific community should welcome a wider discussion on model evaluation, in which subjective judgements made during model development and evaluation are explicitly addressed across various disciplines connected to modelling of physical processes.

The model suitability matrix is flexible framework in the sense that criteria and thresholds used in the decision tree can be modified through stakeholder approach on a case-by-case basis. The framework is implemented using open-source tools and platforms, allowing for relatively easy deployment in most decision-making contexts. However, some components, such as the user interface and the visualization, require further development and tuning. In underpinning upscaling and widespread implementation of forecast-based financing, the model suitability matrix may be an important tool stimulating collaboration between model developers and providers of humanitarian actions. Furthermore, the principles on which the framework builds are not necessarily restricted to the application of forecast-based financing; for any model application, specific criteria may be quantified using similar approaches. As such, the authors urge further development of quantitative and transparent approaches to holistic model evaluation targeted to specific model applications and hereby hope to stimulate a broader scientific discussion that contributes to enhancing the practical value of models in the context of decision-making, management and disaster risk reduction.

Data and software availability

Globally available data used in this study is listed in Table 2.1 with relevant references. Local data was kindly provided by *The Red Cross Red Crescent Climate Centre* and partners associated with implementation of forecast-based financing in Togo. All models constructed in this study are based on open-source code: wflow_sbm is available from GitHub (https://github.com/openstreams/wflow), and the machine learning models and hybrid models were implemented in Python using the sklearn [51] and Keras [55] libraries. Source-codes of these models are available upon request.

CRediT authorship contribution statement

Jenny Sjåstad Hagen: Conceptualization, Methodology, Investigation, Validation, Formal analysis, Visualization, Software, Writing - original draft, Writing - review & editing. **Andrew Cutler:** Methodology, Investigation, Validation, Software, Writing - original draft. **Patricia Trambauer:** Methodology, Resources, Supervision, Conceptualization, Writing - review & editing. **Albrecht Weerts:** Methodology, Resources, Conceptualization, Software, Data curation, Supervision. **Pablo Suarez:** Methodology, Resources, Conceptualization, Supervision. **Dimitri Solomatine:** Methodology, Resources, Conceptualization, Supervision.

Declaration of competing interest

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

Acknowledgements

This research was carried out as part of separate projects at *Deltares* and the *Red Cross Red Crescent Climate Centre*, with funding from the Global Facility for Disaster Reduction and Recovery. The research did not receive any specific grant from funding agencies in the public, commercial or non-profit sectors. The authors would like to thank Janot Mendler de Suarez, Herman Dolder and Eugene Kombaté Nawanti for valuable contributions to the research. The authors would also like to thank *Togo Red Cross*, *Nangbéto Hydropower Dam*, *Specialized Environmental Modeling* and *Agence Nationale de la Protection Civile du Togo* for sharing of data and information regarding FUNES.

References

[1] EM-DAT. The emergency events database. Retrieved from . www.emdat.be; 2018.
[2] A6 IPCC. Climate Change 2014: Synthesis Report. In: Core Writing Team, Pachauri RK, Meyer LA, editors. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, Switzerland: IPCC; 2014.
[3] Kundzewicz ZW, et al. Flood risk and climate change: global and regional perspectives. Hydrol Sci J. 2014;59(1):1–28. https://doi.org/10.1080/02626667.2013.857411.
[4] Arnell NW, Gosling SN. The impacts of climate change on river flood risk at the global scale. Clim Change. 2016;134(3):387–401. https://doi.org/10.1007/s10584-014-1084-5.
[5] Tanoue M, Hirabayashi Y, Ikeuchi H. Global-scale river flood vulnerability in the last 50 years. Sci Rep. 2016;6(1):36021. https://doi.org/10.1038/srep36021.
[6] Van Aalst M, Kellett J, Pichon F, Mitchell T. Incentives in Disaster Risk Management and Humanitarian Response. Background note for the World Development Report 2014. Washington DC: The World Bank; 2013.
[7] Ehret U, Götzinger J, Bárdossy A, Pegram GGS. Radar-based flood forecasting in small catchments, exemplified by the Goldersbach catchment, Germany. International Journal of River Basin Management. 2008;6(4):323–9. https://doi.org/10.1080/15715124.2008.9635359.
[8] Werner M, Cranston M, Harrison T, Whitfield D, Schellekens J. Recent developments in operational flood forecasting in England, Wales and Scotland. Meteorological Applications. 2009;16(1):13–22. https://doi.org/10.1002/met.124.
[9] Pappenberger F, Thielen J, Del Medico M. The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System. Hydrol Process. 2011;25(7):1091–113. https://doi.org/10.1002/hyp.7772.
[10] Burek PA, Dutra E, Alfieri L, Burek P, Dutra E, Krzeminski B, et al. GloFAS–global ensemble streamflow forecasting and flood early warning. Hydrol Earth Syst Sci. 2013;17:1161–75. https://doi.org/10.5194/hessd-9-12293-2012.
[11] Coughlan de Perez E, van den Hurk B, van Aalst MK, Jongman B, Klose T, Suarez P. Forecast-based financing: an approach for catalyzing humanitarian action based on extreme weather and climate forecasts. Natural Hazards and Earth System Science. 2015;15(4):895–904. https://doi.org/10.5194/nhess-15-895-2015.
[12] Wilkinson E, Weingärtner L, Choularton R, Bailey M, Todd M, Kniveton D, et al. Implementing forecast-based early action at scale. Retrieved from . http://lib.riskreductionafrica.org/bitstream/handle/123456789/1501/Forecastinghazards, adverting disastersImplementingforecast-basedearlyactionatscale.pdf?sequence=1; 2018.
[13] Amoussou E. Analyse hydrométéorologique des crues dans le bassin-versant du Mono en Afrique de l'Ouest avec un modèle conceptuel pluie- débit. Fondation Maison Des Sciences de l'homme, FMSH-WP-20; 2015. p. 1–27 Retrieved from . http://www.fmsh.fr.
[14] Ntajal J, Lamptey BL, Sogbedji JM, Kpotivi W-BK, Joshua A. Rainfall trends and flood frequency analyses in the lower Mono River basin in Togo, West Africa. International Journal of Advance Research, IJOAROrg International Journal of Advance Research. 2016;4(10):2320–9186 Retrieved from . http://www.ijoar.org.
[15] Hobson EL. Mapping & assessment of clean energy mini-grid experiences in West-Africa. Oldenburg, Germany. Retrieved from . http://www.ecreee.org/sites/default/files/mapping_and_assessment_of_existing_clean_energy_mini-grid_experiences_in_west_africa_ecreee.pdf; 2016.
[16] Sultan B, Labadi K, Guégan J-F, Janicot S. Climate drives the meningitis epidemics onset in West Africa. PLoS Med. 2005;2(1):43–9. https://doi.org/10.1371/journal.pmed.0020006.
[17] Gu G, Adler RF. Seasonal evolution and variability associated with the West African monsoon system. J Climate. 2004;17(17):3364–77. https://doi.org/10.1175/1520-0442(2004)017<3364:SEAVAW>2.0.CO;2.
[18] Ongoma V, Batebana K, Ogwang BA, Sein ZMM, Ogou FK, Ngarukiyimana JP. Rainfall characteristics over Togo and their related atmospheric circulation anomalies. Journal of Environmental and Agricultural Sciences JEAS. 2014;5:34–48 Retrieved from . http://41.89.55.71:8080/xmlui/handle/123456789/1756.
[19] Amoussou E. Variabilité pluviométrique et dynamique hydro-sédimentaire du bassin versant du complexe fluvio-lagunaire Mono-Ahémé-Couffo (Afrique de l'ouest). Retrieved from . https://hal.archives-ouvertes.fr/tel-00493898/; 2010.
[20] Amoussou E, Tramblay Y, Totin HSV, Mahé G, Camberlin P. Dynamique et modélisation des crues dans le bassin du Mono à Nangbéto (Togo/Bénin) Dynamics and modelling of floods in the river basin of Mono in Nangbeto, Togo/Benin. Hydrological Sciences Journal – Journal Des Sciences Hydrologiques. 2014;59(11). https://doi.org/10.1080/02626667.2013.871015.
[21] Djaman K, Sharma V, Rudnick DR, Koudahe K, Irmak S, Adambounou Amouzou K, et al. Spatial and temporal variation in precipitation in Togo. International Journal of Hydrology. 2017;1(4):97–105. https://doi.org/10.15406/ijh.2017.01.00019.
[22] Tramblay Y, Amoussou E, Dorigo W, Mahé G. Flood risk under future climate in data sparse regions: linking extreme value models and flood generating processes. J Hydrol. 2014;519:549–58. https://doi.org/10.1016/J.JHYDROL.2014.07.052.
[23] Dolder HG. A method for using pre-computed scenarios of physically-based spatially-distributed hydrologic models in flood forecasting systems. Brigham Young University; 2015 Retrieved from . https://scholarsarchive.byu.edu/etd/5676.

[24] Fatichi S, Vivoni ER, Ogden FL, Ivanov VY, Mirus B, Gochis D, et al. An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. J Hydrol. 2016;537:45–60. https://doi.org/10.1016/J.JHYDROL.2016.03.026.
[25] Mosavi A, Ozturk P, Chau K, Mosavi A, Ozturk P, Chau K. Flood prediction using machine learning models: literature review. Water. 2018;10(11):1536. https://doi.org/10.3390/w10111536.
[26] Beck HE, van Dijk AIJM, Levizzani V, Schellekens J, Miralles DG, Martens B, et al. MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. Hydrol Earth Syst Sci. 2017;21(1):589–615. https://doi.org/10.5194/hess-21-589-2017.
[27] Dinku T, Chidzambwa S, Ceccato P, Connor SJ, Ropelewski CF. Validation of high-resolution satellite rainfall products over complex terrain. International Journal of Remote Sensing. 2008;29(14):4097–110. https://doi.org/10.1080/01431160701772526.
[28] Re3data.org. EartH2Observe; 2018. https://doi.org/10.17616/R32K9P.
[29] Brönnimann S. Weather extremes in an ensemble of historical reanalyses. In: Brönnimann S, editor. Historical weather extremes in reanalysesGeographica Bernensia G92; 2017. p. 7–22. https://doi.org/10.4480/GB2017.G92.01.
[30] Molteni F, Buizza R, Palmer TN, Petroliagis T. The ECMWF ensemble prediction system: methodology and validation. Q J Roy Meteorol Soc. 1996;122(529):73–119. https://doi.org/10.1002/qj.49712252905.
[31] Imhoff RO, van Verseveld WJ, van Osnabrugge B, Weerts AH. Scaling point-scale (pedo) transfer functions to seamless large-domain parameter estimates for high-resolution distributed hydrologic modeling: an example for the Rhine river. Water Resour Res. 2020;56. https://doi.org/10.1029/2019WR026807.
[32] Broxton PD, Zeng X, Sulla-Menashe D, Troch PA. A global land cover climatology using MODIS data. Journal of Applied Meteorology and Climatology. 2014;53(6):1593–605. https://doi.org/10.1175/JAMC-D-13-0270.1.
[33] Nachtergaele F, Van Velthuizen H, Verelst L, Batjes CN, Dijkshoorn K, Van Engelen V, et al. Harmonized world soil database - version 1.1. Laxenburg. Retrieved from . http://www.fao.org/3/a-aq361e.pdf; 2009.
[34] Lehner B, Grill G. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrol Process. 2013;27(15):2171–86. https://doi.org/10.1002/hyp.9740.
[35] Jarvis A, Rubiano J, Nelson A, Farrow A, Mulligan M. Practical use of SRTM data in the tropics – comparisons with digital elevation models generated from cartographic data. Cali, Colombia. Retrieved from . http://ciat-library.ciat.cgiar.org/articulos_ciat/Jarvis4.pdf; 2004.
[36] Zwart SJ, Mishra B, Dembélé M. Satellite rainfall for food security on the African continent: performance and accuracy of seven rainfall products between 2001 and 2016. Retrieved from . https://research.utwente.nl/en/publications/satellite-rainfall-for-food-security-on-the-african-continent-per; 2018, May 9.
[37] Schellekens J, Becker BPJ, Donchyts G, Goorden N, Hoogewoud JC, Patzke S, et al. OpenStreams: open source components as building blocks for integrated hydrological models. EGU general assembly 2012, held 22–27 April, 2012 in Vienna, Austria, vol. 14. ; 2012. p. 3953 3953. Retrieved from . http://adsabs.harvard.edu/abs/2012EGUGA..14.3953S.
[38] Schellekens J, van Verseveld W, Euser T, Winsemius H, Thiange C, Bouaziz L, et al. openstreams/wflow: unstable-master. Retrieved March 21, 2018, from . https://github.com/openstreams/wflow; 2017.
[39] Lindström G, Johansson B, Persson M, Gardelin M, Bergström S. Development and test of the distributed HBV-96 hydrological model. J Hydrol. 1997;201(1–4):272–88. https://doi.org/10.1016/S0022-1694(97)00041-3.
[40] Edijatno C, Michel C. Un modèle pluie-débit journalier à trois paramètres. La Houille Blanche. 1989;2:113–22. https://doi.org/10.1051/lhb/1989007.
[41] van Dijk A, Peña-Arancibia JL, Wood EF, Sheffield J, Beck HE. Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide. Water Resour Res. 2013;49(5):2729–46. https://doi.org/10.1002/wrcr.20251.
[42] Vertessy RA, Elsenbeer H. Distributed modeling of storm flow generation in an Amazonian rain forest catchment: effects of model parameterization. Water Resour Res. 1999;35(7):2173–87. https://doi.org/10.1029/1999WR900051.
[43] Karssenberg D, Schmitz O, Salamon P, de Jong K, Bierkens MFP. A software framework for construction of process-based stochastic spatio-temporal models and data assimilation. Environ Model Software. 2010;25(4):489–502. https://doi.org/10.1016/J.ENVSOFT.2009.10.004.
[44] Hassaballah K, Mohamed Y, Uhlenbrook S, Biro K. Analysis of streamflow response to land use and land cover changes using satellite data and hydrological modelling: case study of Dinder and Rahad tributaries of the Blue Nile (Ethiopia–Sudan). Hydrol Earth Syst Sci. 2017;21(10):5217–42. https://doi.org/10.5194/hess-21-5217-2017.
[45] López P, Wanders N, Schellekens J, Renzullo LJ, Sutanudjaja EH, Bierkens MFP. Improved large-scale hydrological modelling through the assimilation of streamflow and downscaled satellite soil moisture observations. Hydrol Earth Syst Sci. 2016;20(7):3059–76. https://doi.org/10.5194/hess-20-3059-2016.
[46] Werner M, Schellekens J, Gijsbers P, Van Dijk M, Van Den Akker O, Heynert K. The Delft-FEWS flow forecasting system. Environ Model Software. 2013;40:65–77. https://doi.org/10.1016/j.envsoft.2012.07.010.
[47] Van Dijk A, Bacon D, Barratt D, Crosbie R, Daamen C, Fitch P, et al. Design and development of the Australian Water Resources Assessment system. In: Sims J, Merrin L, Ackland R, Herron N, editors. Water information research and development alliance: science symposium proceedings. Melbourne: CSIRO; 2011. p. 17–27 Retrieved from . https://publications.csiro.au/rpr/pub?list=BRO&pid=csiro:EP116648&sb=RECENT&n=8&rpp=10&page=272&tr=4282&dr=all&dc4.browseYear=2012.
[48] Demargne J, Wu L, Regonda SK, Brown JD, Lee H, He M, et al. The science of NOAA's operational hydrologic ensemble forecast service. Bull Am Meteorol Soc. 2014;95(1):79–98. https://doi.org/10.1175/BAMS-D-12-00081.1.
[49] Weerts AH, Schellekens J, Sperna Weiland F. Real-time geospatial data handling and forecasting: examples from Delft-FEWS forecasting platform/system. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2010;3(3):386–94. https://doi.org/10.1109/JSTARS.2010.2046882.
[50] de Bruin HAR, Trigo IF, Bosveld FC, Meirink JF. A thermodynamically based model for actual evapotranspiration of an extensive grass field close to FAO reference, suitable for remote sensing application. J Hydrometeorol. 2016;17(5):1373–82. https://doi.org/10.1175/JHM-D-15-0006.1.
[51] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011;12:2825–30 Retrieved from . http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.
[52] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006;63(1):3–42.
[53] Chollet F. Xception: deep learning with depthwise separable convolutions. Retrieved from . http://arxiv.org/abs/1610.02357; 2016.
[54] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time-series parsing view project MoDeep view project. The handbook of brain theory and neural networks; 1995 Retrieved from . https://www.researchgate.net/publication/2453996.
[55] Collet F. Keras. GitHub. Retrieved from . https://keras.io/#keras-the-python-deep-learning-library; 2015.
[56] Abrahart RJ, Anctil F, Coulibaly P, Dawson CW, Mount NJ, See LM, et al. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. Progress in Physical Geography. 2012;36(4):480–513. https://doi.org/10.1177/0309133312444943.
[57] Solomatine DP, Ostfeld A. Data-driven modelling: some past experiences and new approaches. Journal of Hydroinformatics. 2008;10(1):3. https://doi.org/10.2166/hydro.2008.015.
[58] Beven K. How far can we go in distributed hydrological modelling? Dalton lecture: how far can we go in distributed hydrological modelling? How far can we go in distributed hydrological modelling? The Dalton lecture. Hydrol Earth Syst Sci. 2001;5(51):1–12 Retrieved from . https://hal.archives-ouvertes.fr/file/index/docid/304564/filename/hess-5-1-2001.pdf.
[59] Han S, Pool J, Tran J, Dally WJ. Learning both weights and connections for efficient neural networks. Adv Neural Inf Proces Syst. 2015;1:1135–43.
[60] Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research. 2014;15 Retrieved from . http://www.machinelearning.ru/wiki/images/f/f2/MMP-Dropout_full_article.pdf.
[61] Ritter A, Muñoz-Carpena R. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J Hydrol. 2013;480:33–45. https://doi.org/10.1016/J.JHYDROL.2012.12.004.
[62] Krause P, Boyle DP, Bäse F. Comparison of different efficiency criteria for hydrological model assessment. . 2005;5:89–97 Retrieved from . https://www.adv-geosci.net/5/89/2005/adgeo-5-89-2005.pdf.
[63] Brown JD, Demargne J, Seo D-J, Liu Y. The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environ Model Software. 2010;25(7):854–72. https://doi.org/10.1016/J.ENVSOFT.2010.01.009.
[64] Cloke HL, Pappenberger F. Ensemble flood forecasting: a review. J Hydrol. 2009;375(3–4):613–26. https://doi.org/10.1016/J.JHYDROL.2009.06.005.