Building a Multimodel Flood Prediction System with the TIGGE Archive

ERVIN ZSÓTÉR,^{a,d} FLORIAN PAPPENBERGER,^{a,b} PAUL SMITH,^{a,c} REBECCA ELIZABETH EMERTON,^{a,d}
EMANUEL DUTRA,^a FREDRIK WETTERHALL,^a DAVID RICHARDSON,^a KONRAD BOGNER,^{a,c} AND
GIANPAOLO BALSAMO^a

^a European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

^b College of Hydrology and Water Resources, Hohai University, Nanjing, China

^c Lancaster Environment Centre, Lancaster University, Lancaster, United Kingdom

^d Department of Geography and Environmental Science, University of Reading, Reading, United Kingdom

^e Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

(Manuscript received 28 July 2015, in final form 23 July 2016)

ABSTRACT

In the last decade operational probabilistic ensemble flood forecasts have become common in supporting decision-making processes leading to risk reduction. Ensemble forecasts can assess uncertainty, but they are limited to the uncertainty in a specific modeling system. Many of the current operational flood prediction systems use a multimodel approach to better represent the uncertainty arising from insufficient model structure. This study presents a multimodel approach to building a global flood prediction system using multiple atmospheric reanalysis datasets for river initial conditions and multiple TIGGE forcing inputs to the ECMWF land surface model. A sensitivity study is carried out to clarify the effect of using archive ensemble meteorological predictions and uncoupled land surface models. The probabilistic discharge forecasts derived from the different atmospheric models are compared with those from the multimodel combination. The potential for further improving forecast skill by bias correction and Bayesian model averaging is examined. The results show that the impact of the different TIGGE input variables in the HTESSEL/Catchment-Based Macroscale Floodplain model (CaMa-Flood) setup is rather limited other than for precipitation. This provides a sufficient basis for evaluation of the multimodel discharge predictions. The results also highlight that the three applied reanalysis datasets have different error characteristics that allow for large potential gains with a multimodel combination. It is shown that large improvements to the forecast performance for all models can be achieved through appropriate statistical postprocessing (bias and spread correction). A simple multimodel combination generally improves the forecasts, while a more advanced combination using Bayesian model averaging provides further benefits.

1. Introduction

Operational probabilistic ensemble flood forecasts have become more common in the last decade (Cloke and Pappenberger 2009; Demargne et al. 2014; Olsson and Lindström 2008). Ensemble forecasts are a good way of assessing forecast uncertainty, but they are limited to the uncertainty captured by a specific modeling system. A multimodel approach can address

this shortcoming and provide a more complete representation of the uncertainty in the model structure, also potentially reducing the errors (Krishnamurti et al. 1999).

“Multimodel” can refer to systems using multiple meteorological models, hydrological models, or both (Velázquez et al. 2011). According to Emerton et al. (2016), among the many regional-scale operational hydrological ensemble prediction systems across the globe, at present there are six large-scale (continental and global) models: four that run at continental scale over Europe, Australia, and the United States and two that are available globally. The U.S. Hydrologic Ensemble Forecast Service (HEFS), run by the National Weather Service (NWS; Demargne et al. 2014), and the Global Flood Forecasting Information System (GLOFFIS), a

Denotes Open Access content.

Corresponding author address: E. Zsótér, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, United Kingdom.
E-mail: ervin.zsoter@ecmwf.int

DOI: 10.1175/JHM-D-15-0130.1

recent development at Deltares in the Netherlands, are examples of systems using different hydrological models as well as multiple meteorological inputs. The European Flood Awareness System (EFAS) developed by the Joint Research Centre (JRC) of the European Commission and ECMWF operates using a single hydrological model with multimodel meteorological input (Thielen et al. 2009). Finally, the European Hydrological Predictions for the Environment (E-HYPE) Water in Europe Today (WET) model of the Swedish Meteorological and Hydrological Institute (SMHI; Donnelly et al. 2015), the Australian Flood Forecasting and Warning Service, and the Global Flood Awareness System (GloFAS; Alfieri et al. 2013), running in collaboration between ECMWF and JRC, all use one main hydrological model and one meteorological model input.

While the multimodel approach has traditionally involved the use of multiple forcing inputs and hydrological models to generate discharge forecasts, it also allows for consideration of multiple initial conditions. In keeping with GloFAS, this paper uses atmospheric reanalysis data to generate the initial conditions of the land surface components of the forecasting system; therefore, a multimodel approach based on three reanalysis datasets is trialed.

The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE; Bougeault et al. 2010) archive is an invaluable source of multimodel meteorological forcing data. The archive has attracted attention among hydrological forecasters and is already being extensively used in hydrological applications. The first published example of a hydrometeorological forecasting application was by Pappenberger et al. (2008). In that paper, the forecasts of nine TIGGE centers were used within the setting of EFAS for a case study of a flood event in Romania in October 2007 and showed that the lead time of flood warnings could be improved by up to 4 days through the use of multiple forecasting models rather than a single model. This study and other subsequent studies using TIGGE multimodel data (e.g., He et al. 2009, 2010; Bao and Zhao 2012) have indicated that combining different models not only increases the skill, but also the lead time at which warnings could be issued. He et al. (2009) highlighted this and further showed that individual systems of the multimodel forecast have systematic errors in time and space that would require temporal and spatial postprocessing. Such postprocessing should carefully maintain spatial, temporal, and intervariable correlations; otherwise, they lead to deteriorating hydrological forecast skill.

The scientific literature contains numerous studies on methods that can lead to significant gain in forecast skill by combining and postprocessing different forecast systems. Statistical ensemble postprocessing techniques target the generation of sharp and reliable probabilistic forecasts from ensemble outputs. Hagedorn et al. (2012) showed, based on TIGGE, that by considering an equal-weight multimodel approach, a selection of best NWP models might be needed to gain skill on the best-performing single model. In addition to this, the calibration of the best single model using a reforecast dataset can lead to comparable or even superior quality to the multimodel prediction. Gneiting and Katzfuss (2014) focus on various methodologies that require weighting of the different contributing forecasts to optimize model error corrections. They recommend the application of well-established techniques in the operational environment such as the nonhomogeneous regression or Bayesian model averaging (BMA). The BMA method generates calibrated and sharp probability density functions (PDFs) from ensemble forecasts (Raftery et al. 2005), where the predictive PDF is a weighted average of the PDFs centered on the bias-corrected forecasts. The weights reflect the relative skill of the individual members over a training period. The BMA has been widely used and proved to be beneficial in hydrological ensemble systems (e.g., Ajami et al. 2007; Cane et al. 2013; Dong et al. 2013; Liang et al. 2013; Todini 2008; Vrugt and Robinson 2007).

Previous studies have used hydrological models, rather than land surface models, to analyze the benefits of multimodel forecasting and have focused on individual catchments. The potential of multimodel forecasts at the regional or continental scale shown in previous studies provides the motivation for building a global multimodel hydrometeorological forecasting system.

In this study we present our experiences in building a multimodel hydrometeorological forecasting system. Global ensemble discharge forecasts with a 10-day horizon are generated using the ECMWF land surface model and a river-routing model. The multimodel approach arises from the use of meteorological forecasts from four models in the TIGGE archive and the derivation of river initial conditions using three global reanalysis datasets. The main focus of our study is the quality of the discharge forecasts derived from the TIGGE data. We analyze the Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land (HTESSEL)/Catchment-Based Macroscale Floodplain model (CaMa-Flood) setup and the scope for error reduction by applying the multimodel approach and different postprocessing methods on the forecast data. Three sets of experiments are undertaken to test

(i) the sensitivity of the forecasting system to the input variables, (ii) the potential improvements in forecasting historical discharge that can be achieved by a combination of different reanalysis datasets, and (iii) the use of bias correction and model combination to improve the predictive distribution of the forecasts.

In [section 2](#) the datasets, models, and methodology used throughout the paper are described. [Section 3](#) summarizes the discharge experiments we produced and analyzed. In [section 4](#), we provide the results, while [section 5](#) gives conclusions to the paper.

2. System description and datasets

a. HTESSEL land surface model

The hydrological component of this study was the HTESSEL ([Balsamo et al. 2009, 2011](#)) land surface model. The HTESSEL scheme follows a mosaic (or tiling) approach where the grid boxes are divided into patches (or tiles), with up to six fractions over land (bare ground, low and high vegetation, intercepted water, and shaded and exposed snow) and two extra tiles over water (open and frozen water) exchanging energy and water with the atmosphere. The model is part of the Integrated Forecast System (IFS) at ECMWF and is used in coupled atmosphere–surface mode on time ranges from medium range to seasonal forecasts. In addition, the model provides a research test bed for applications where the land surface model can run in a stand-alone mode. In this so-called “offline” version the model is forced with near-surface meteorological input (temperature, specific humidity, wind speed, and surface pressure), radiative fluxes (downward solar and thermal radiation), and water fluxes (liquid and solid precipitation). This offline methodology has been explored in various research applications where HTESSEL or other models were applied (e.g., [Agustí-Panareda et al. 2010](#); [Dutra et al. 2011](#); [Haddeland et al. 2011](#)).

b. CaMa-Flood river routing

CaMa-Flood ([Yamazaki et al. 2011](#)) was used to integrate HTESSEL runoff over the river network into discharge. CaMa-Flood is a distributed global river-routing model that routes runoff to oceans or inland seas using a river network map. A major advantage of CaMa-Flood is the explicit representation of water level and flooded area in addition to river discharge. The relationship between water storage (the only prognostic variable), water level, and flooded area is determined on the basis of the subgrid-scale topographic parameters based on a 1-km digital elevation model.

c. TIGGE forecasts

The atmospheric forcing for the forecast experiments is taken from the TIGGE archive where all variables are available on a standard 6-h forecast frequency. The ensemble systems of ECMWF, the Met Office (UKMO), the National Centers for Environmental Prediction (NCEP), and the China Meteorological Administration (CMA) provide, in the TIGGE archive, meteorological forcing fields from the 0000 UTC runs with 6-h frequency starting from 2006 to 2008 depending on the model. All four models were only available with the complete forcing variable set from August 2008. ECMWF was available with 50 ensemble members on 32-km horizontal resolution (~50 km before January 2010) up to 15 days ahead, UKMO was available with 23 members on ~60-km horizontal resolution (~90 km before March 2010) also up to 15 days ahead, NCEP was available with 20 members on ~110-km horizontal resolution up to 16 days ahead, and finally CMA was available with 14 members on ~60-km horizontal resolution up to 10 days ahead. In testing the sensitivity of the experimental setup to meteorological forcing (see [section 4a](#)) the ECMWF control forecasts were used, extracted directly from ECMWF’s Meteorological Archival and Retrieval System (MARS), where the meteorological variables are available without the TIGGE restrictions. These have the same resolution as the 50 ensemble members but start from the unperturbed analysis.

d. Reanalysis data

The discharge modeling experiments require reanalysis data, which are used to provide the climate and the initial conditions needed for the HTESSEL land surface model runs and to produce the river initial conditions required in the CaMa-Flood routing part of the TIGGE forecast experiments.

In this study we have used three different reanalysis datasets: two produced by ECMWF, ERA-Interim (hereafter ERAI) and ERA-Interim/Land with Global Precipitation Climatology Project, version 2.2 (GPCP v2.2), precipitation ([Huffman et al. 2009](#)) correction (hereafter ERAI-Land; [Balsamo et al. 2015](#)), and a third, the Modern-Era Retrospective Analysis for Research and Applications (MERRA) land upgrade (MERRA-Land) produced by NOAA. The combination of these three sources was a proof of concept to potential added value of the multi-initial conditions.

ERAI is ECMWF’s global atmospheric reanalysis from 1979 to present produced with an older (2006) version of the ECMWF IFS on a T255 spectral resolution ([Dee et al. 2011](#)). ERAI-Land is a version of ERAI

FIG. 1. Location of discharge observing stations that could be processed in the discharge experiments. The blue points are used in both the reanalysis (at least 15 years of data available in 1981–2010) and in the TIGGE forecast experiment (at least 80% of days available from August 2008 to May 2010) evaluation. The yellow points provide enough observation only for the reanalysis while the red points have enough data available only for the TIGGE forecasts.

at the same 80-km spatial resolution with improvements for land surface. It was produced in offline mode with a 2014 version of the HTESSEL land surface model using atmospheric forcing from ERAI, with precipitation adjustments based on GPCP v2.2, where the ERAI 3-hourly precipitation is rescaled to match the monthly accumulated precipitation provided by the GPCP v2.2 product [for more details, please consult [Balsamo et al. \(2010\)](#)].

The MERRA-Land dataset is similar to ERAI-Land in that it is a land-only version of the MERRA land model component, produced also in offline mode, using improved precipitation forcing and an improved version of the catchment land surface model ([Reichle et al. 2011](#)).

e. Discharge data

In this study a subset of the observations available in GloFAS was used, mainly originating from the Global Runoff Data Centre (GRDC) archive. The GRDC is the digital worldwide depository of discharge data and associated metadata. It is an international archive of data started in 1812, and it fosters multinational and global long-term hydrological studies.

For the discharge modeling, a dataset of 1121 stations with upstream areas over 10 000 km² was available until the end of 2013. GRDC has a gradually decreasing number of stations with data in the archive limiting their use for more recent years. For the forecast discharge, we limited our analyses to the period from August 2008 to May 2010. This period provided the optimal compromise in increasing the sample size between the length of the

period and the number of stations with good data coverage. For the reanalysis discharge experiments and also for generating the observed discharge climate, stations with a minimum of 15 years of available observations were used in the 30-yr period from 1981 to 2010. For the forecast experiments, stations with at least 80% of the observations available were used in the 22-month period from August 2008 to May 2010. [Figure 1](#) shows the observation availability in the reanalysis and TIGGE forecast experiments. It highlights that for the reanalysis the coverage is better globally, with about 850 stations, while the forecast experiments have around 550 stations with large missing areas, mainly in Africa and Asia.

f. Forecasting system setup

To produce runoff from the TIGGE atmospheric ensemble variables (see [section 2c](#)), HTESSEL experiments were run with 6-hourly forcing frequency and hourly model time step. For the instantaneous variables (such as 2-m temperature), linear interpolation was used to move from the 6-h to hourly time step used in the HTESSEL simulations. For accumulated variables (such as precipitation), a disaggregation algorithm that conserves the 6-hourly totals was used. The disaggregation algorithm divides into hourly values based on a linear combination of the current and adjacent 6-hourly totals with weights derived from the time differences.

The climate and the initial conditions needed for the HTESSEL land surface model runs to produce runoff were taken from ERAI-Land, the same initial conditions for all models and ensemble members without

perturbations. The other two reanalysis datasets could also be used to initialize HTESSEL, but the variability on the resulting TIGGE runoff (and thus on the TIGGE discharge) would be very small compared with the impact of the TIGGE atmospheric forcing (especially precipitation, see also [section 4a](#)) and the impact of the TIGGE forecast routing initialization (see [section 3b](#) for further details).

HTESSEL was set to T255 spectral resolution (~ 80 km). This was the horizontal resolution used in ERAI and was an adequate compromise between the highest (ECMWF mainly ~ 50 km) and lowest (NCEP with ~ 110 km) forcing model resolution that also allowed fast enough computations. The TIGGE forcing fields were transformed to T255 using bilinear interpolation.

The TIGGE archive includes variables at the surface and several pressure levels. However, variables are not available on model levels, and as such, temperature, wind, and humidity at the surface (i.e., 2 m for temperature and humidity and 10 m for wind) were used in HTESSEL rather than on the preferred lowest model level (LML).

Similarly, TIGGE contains several radiation variables, but not the downward radiations required by HTESSEL. To run HTESSEL without major technical modifications, we had to use a radiation replacement for all TIGGE models and ensemble members. We used ERAI-Land for this purpose, as it does not favor any of the TIGGE models used in this study. This way, for one daily run the same single radiation forecast was used for all ensemble members and all models. These 10-day radiation forecasts were built from 12-h ERAI-Land short-range predictions. To reduce the possible spinup effects in the first hours of the ERAI-Land forecasts, the 6–18-h radiation fluxes were combined (as 12-h sections) from subsequent 0000 and 1200 UTC runs, following the approach described in [Balsamo et al. \(2015\)](#). The sensitivity to the HTESSEL input variables will be discussed in [section 4a](#).

In this study we were able to process four models out of the 10 global models archived in TIGGE: ECMWF, UKMO, NCEP, and CMA. The other six models do not archive one or more of the forcing variables, in addition to the downward radiation, required for this study.

The runoff produced by HTESSEL for TIGGE was routed over the river network by CaMa-Flood. These relatively short experiments for the TIGGE forecasts required initial river conditions. These were provided by three CaMa-Flood runs for the 1980–2010 period with ERAI, ERAI-Land, and MERRA-Land runoff input.

The discharge forecasts were produced by CaMa-Flood out to 10 days ($T + 240$ h), the longest forecast horizon common to all models. No perturbations were

applied on the river initial conditions for the ensemble members. The forecasts were extracted from the CaMa-Flood 15-arc-min (~ 25 km) model grid for every 24-h similarly to the 24-h reporting frequency of the discharge observations.

3. Experiments

The main focus of the experiments was on the quality of the discharge forecasts derived from the TIGGE data. Three sets of experiments were performed to test the HTESSEL/CaMa-Flood setup and the scope for error reduction by applying the multimodel approach and different postprocessing methods:

- Discharge sensitivity to meteorological forcing: The first experiment ([section 4a](#)) tests the sensitivity of the forecasting system to the input variables.
- Reanalysis impact on discharge: The second experiment ([section 4b](#)) evaluates the potential improvements on the historical discharge that can be achieved by a combination of different reanalysis datasets.
- Improving the forecast distribution: In the third experiment ([section 4c](#)), the use of bias correction and model combination to improve the predictive distribution of the forecast is considered.

a. Discharge sensitivity to meteorological forcing

In [section 2f](#), a number of compromises in the coupling of HTESSEL and forecasts from the TIGGE archive were introduced. Sensitivity experiments were conducted to study the impact of these. [Table 1](#) provides a short description of the experiments.

The baseline for the comparisons is the discharge forecasts generated by HTESSEL and CaMa-Flood driven by ECMWF ensemble control (EC) forecasts. These forecasts were produced weekly (at 0000 UTC) throughout 2008–12 to cover several seasons (~ 260 forecast runs in total). In the baseline setup, the LML meteorological output for temperature, wind, and humidity was used to drive HTESSEL.

The first sensitivity test (Surf vs LML) was to replace these LML values with the surface values (as 2-m temperature and humidity and 10-m wind) from the same model run. This mirrors the change needed to make use of the TIGGE archive. Because of limitations in the TIGGE archive, the ERAI-Land radiation was used for all forecasts. Substitution of the ECMWF EC radiation in the HTESSEL input by ERAI-Land is the second sensitivity test (Rad). Further to this, substitution of the wind (Wind), temperature, humidity, and surface pressure together (THP), and precipitation (Prec) from ERAI-Land in place of the ECMWF EC run values was also

TABLE 1. Description of the sensitivity experiments with the ECMWF EC forecasts. The baseline is the reference run at the LML for wind, temperature, and humidity forcing. The other experiments are with different changes for the forcing variables. First, the LML is changed to surface (Surf), then different variables of the EC and their combinations are substituted by ERAI-Land data. Roman font means EC forcing input while italicized font denotes substituted ERAI-Land input.

Sensitivity expt	Forcing variable setup			
	Rad	THP	Wind	Prec
Baseline	EC-Surf	EC-LML	EC-LML	EC-Surf
Surf vs LML	EC-Surf	EC-Surf	EC-Surf	EC-Surf
Rad	<i>ERAI-Surf</i>	EC-LML	EC-LML	EC-Surf
Wind	EC-Surf	EC-LML	<i>ERAI-LML</i>	EC-Surf
THP	EC-Surf	<i>ERAI-LML</i>	EC-LML	EC-Surf
Rad + THP	<i>ERAI-Surf</i>	<i>ERAI-LML</i>	EC-LML	EC-Surf
Rad + THP + Wind	<i>ERAI-Surf</i>	<i>ERAI-LML</i>	<i>ERAI-LML</i>	EC-Surf
Rad + THP + Wind + Prec	<i>ERAI-Surf</i>	<i>ERAI-LML</i>	<i>ERAI-LML</i>	<i>ERAI-Surf</i>

evaluated. Temperature and humidity were analyzed together because of the sensitive nature of the balance between these two variables. Although these changes were not applied on the TIGGE data, they give a more complete picture on sensitivity to the forcing variables. This puts into context the discharge errors that we indirectly introduced through the TIGGE-HTESSEL setup changes.

The impact on the errors was compared by evaluating the ratio of the magnitude (absolute value) of the discharges to the baseline experiments, discharge value. These changes in relative discharge were computed for each station as the average of the relative changes over all runs (in the 2008–12 period with weekly runs) and also as a global average of all available stations.

b. Reanalysis impact on discharge

For the forecast of CaMa-Flood routing, the river initial conditions are provided by reanalysis-based simulations (see section 2d). They do not make use of observed river flow and therefore are an estimate of the observed values. The quality of the forecast discharge is expected to be strongly dependent on the skill of this reanalysis-derived historical discharge. This is highlighted in Fig. 2, where ERAI-Land, ERAI, and MERRA-Land are compared for a station in the United States for a 4-yr period.

Each of these reanalyses provides different error characteristics that can potentially be harnessed by using a multimodel approach. For this station ERAI has a tendency to produce occasional high peaks, while MERRA-Land has a strong negative bias. Although Fig. 2 is only a single example, it highlights the large variability between these reanalysis datasets and therefore a potentially severe underestimation of the uncertainties in the subsequent forecast experiments by using only a single initialization dataset.

The impact of the multimodel approach was analyzed by experiments with the historical discharges derived

from ERAI, ERAI-Land, and MERRA-Land inputs. Three sets of CaMa-Flood routing runs were performed for each of the four TIGGE models for the whole 22-month period in 2008–10, each initialized from one of the three reanalysis-derived historical river conditions. The performance of the historical discharge was evaluated independently of the TIGGE forecasts on the period of 1981–2010.

c. Improving the forecast distribution

In the third group of experiments a number of post-processing techniques were applied at each site with the aim of improving the forecast distribution for the observed data. Here we outline the techniques with reference to a single site and forecast origin t . The forecast values available are denoted $f_{m,j,t,i}$, where m indices over the forecast products (ECMWF, UKMO, NCEP, and CMA), j indices over the N_m ensemble members in forecast product m , and $i = 1, \dots, 10$ indicates the available lead times.

FIG. 2. Example of discharge produced by ERAI-Land (red), ERAI (green), and MERRA-Land (blue) forcing and the corresponding observations (black) for a GRDC station on the Rainy River at Manitou Rapids in the United States.

1) BIAS CORRECTION

As a first step we analyzed the biases of the data. As described in [section 3b](#), the historical river initial conditions have potentially large errors. In addition, the variability of the discharge in a 10-day forecast horizon is generally much smaller than derived from reanalysis over a long period. Therefore, any timing or magnitude error in the historical discharge provided initial conditions means the forecast errors can be very large and will change only slightly, in relative terms, throughout the 10-day forecast period.

As bias was expected to be a very important aspect of the errors, three methods of computing the bias correction $e_{m,j,t,i}$ to add to the forecast $f_{m,j,t,i}$ were proposed. The first of these is to apply no correction (or uncorrected); that is, $e_{m,j,t,i} = 0$ in all cases. The second method, referred to as 30-day correction, removes the mean bias of the 30-day period preceding the actual forecast run for each forecast product at each specified forecast range. The mean bias is computed as an average error of the ensemble mean over a 30-day period. In this case, given a series of 30 dates $t = 1, \dots, N_{30}$ and observed discharge data y_t , the bias corrections are given by

$$\frac{\sum_{t=1}^{N_{30}} \sum_{j=1}^{N_m} (y_t - f_{m,j,t-i})}{N_{30} N_m}.$$

The third correction method, referred to as initial time correction, focused specifically on the historical discharge-based initial condition errors. The error at initialization of the routing $f_{m,j,t,0}$, that is, the error of the historical discharge, was used as a correction for all forecast ranges. This initial time correction gives

$$e_{m,j,t,i} = y_t - f_{m,j,t,0}.$$

This method therefore uses a specific error correction for each individual forecast run from day 1 to day 10. Because of the common initialization, the initial time correction was the same for all four TIGGE models for all three historical discharge experiments, respectively.

2) MULTIMODEL COMBINATION

To investigate the potential further benefits of combining different forecast products, two model combination strategies were trialed. The naïve combination strategy [also referred to as multimodel combination (MM)] was based on utilizing a grand ensemble with each member having equal weight. In this combination, the larger ensembles (the largest being ECMWF with

50 members) get larger weights. In direct analogy to the case of a single forecast product, the cumulative forecast distribution is expressed in terms of the indicator function $\delta(z)$, which takes the value 1 if the statement z is true and 0 otherwise, as

$$\Pr(Y_{t+i} < y) = \frac{\sum_m \sum_{j=1}^{N_m} \delta(f_{m,j,t,i} + e_{m,j,t,i} < y)}{\sum_m N_m}.$$

Here $e_{m,j,t,i}$ indicates one of the three bias corrections we introduced in the previous section.

In the second combination strategy, BMA was used to explore further the effects of weighted combination and a temporally localized bias correction. Since discharge is always positive, the variables were transformed so that their distributions marginalized over time are standard Gaussian. This is achieved using the normal quantile transform ([Krzysztofowicz 1997](#)), with the upper and lower tails handled as in [Coccia and Todini \(2011\)](#). The transformed values of the bias-corrected forecasts and observations are denoted $\tilde{f}_{m,j,t,i}$ and $\tilde{y}_t$, respectively.

This study follows the BMA approach proposed by [Fraley et al. \(2010\)](#) for systems with exchangeable members with the weight $w_{m,t,i}$, linear bias correction (with parameters $a_{m,t,i}$ and $b_{m,t,i}$), and nugget variance $\sigma_{m,t,i}^2$ being identical for each ensemble member within a given forecast product. The resulting cumulative forecast distribution in the transformed space is then a weighted combination of standard Gaussian cumulative distributions Φ , specifically,

$$\Pr(\tilde{Y}_{t+i} < y) = \sum_m w_{m,t,i} \sum_{j=1}^{N_m} \Phi\left(\frac{y - a_{m,t,i} - b_{m,t,i} \tilde{f}_{m,j,t,i}}{\sigma_{m,t,i}}\right).$$

As indicated by the origin and lead time subscripts, the BMA parameters were estimated for each forecast origin and lead time. Estimation bias proceeds by first fitting the linear correction using least squares before estimating the weight and variance terms using maximum likelihood ([Raftery et al. 2005](#)). A moving window of 30 days of data, before the initialization of the forecasts similarly to the 30-day correction, was utilized for the estimation to mimic operational practice.

As the initial conditions were expected to play an important role, a further forecast was introduced in the context of the BMA analysis. The deterministic persistence forecast is, throughout the 10-day forecast range, the most recent observation available at time of issue, that is,

$$f_{m,1,t,i} = y_t.$$

This persistence forecast was also used as a simple reference to compare our forecasts against.

To aid comparison with the naïve combination strategy a similar-sized ensemble of forecasts was generated from the BMA combination by applying ensemble copula coupling (Schefzik et al. 2013) to a sample generated by taking equally spaced quantiles from the forecast distribution and reversing the transformation.

3) VERIFICATION STATISTICS

The forecast distributions were evaluated using the continuous ranked probability score (CRPS; Candille and Talagrand 2005). The CRPS evaluates the global skill of the ensemble prediction systems by measuring a distance between the predicted and the observed cumulative density functions of scalar variables. For a set of dates $t = 1, \dots, N$ with observations and probabilistic forecasts issued with the same lead time (which are realizations of the random variables Y_{t+i}), the CRPS can be defined as

$$\text{CRPS} = \frac{1}{N_t} \sum_{t=1}^{N_t} \int_{-\infty}^{\infty} [\Pr(Y_{t+i} < y) - \delta(y \geq y_t)]^2 dy.$$

The CRPS has a perfect score of 0 and has the advantage of transforming into the mean absolute error for deterministic forecasts and thus providing a simple way of comparing different types of systems. In this study the method of Hersbach (2000) for computing the CRPS from samples was used. The global CRPS reported for each lead time were produced by pooling the samples from all the stations before computing the scores.

As the CRPS has the unit of the physical quantity (e.g., for discharge $\text{m}^3 \text{s}^{-1}$), comparing scores can be problematic and is only meaningful if two homogeneous sample-based scores are compared. For example, different geographical areas or different seasons cannot really be compared. In this study we ensured that, for any comparison of forecast models and postprocessed versions, the samples were homogeneous. We considered the same days in the verification period at each station specifically, and also the same stations in the global analysis, producing equal sample sizes across all compared products.

To help compare results across different stations and areas, we used the CRPS-based skill score (CRPSS) with the reference system of the observed discharge climate in our verification. We produced the daily observed climate for the 30-yr period of 1981–2010 and pooled observations from a 31-day window centered over each day. Observed climate was produced for stations with at

least 10 years of data available in total (310 values) for all days of the year.

Each of the historical discharge experiments produce a time series of discharges (f_t ; $t = 1, \dots, N$), which were compared to the observed data using the mean absolute error (MAE)-based skill score (MAESS) with the observed daily discharge climate (obsclim) as reference,

$$\text{MAESS} = 1 - \frac{\text{MAE}}{\text{MAE}_{\text{obsclim}}},$$

and the sample Pearson correlation coefficient (CORR),

$$\text{CORR} = \frac{\sum_{t=1}^N (f_t - \bar{f})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^N (f_t - \bar{f})^2} \sqrt{\sum_{t=1}^N (y_t - \bar{y})^2}},$$

where the bar denotes the temporal average of the variable. The MAE reflects the ability of the systems to match the actual observed discharge, while the correlation highlights the quality of match between the temporal behavior of the historical forecast time series and the observation time series.

4. Results

First, we present the findings of the sensitivity experiments carried out, using the ECMWF EC forecast, on the impact of the HTESSEL coupling with the TIGGE meteorological input. Then we compare the quality of the historical discharge produced from the ERAI, ERAI-Land, and MERRA-Land datasets and the impact of their combination. Finally, from the large number of forecast products described in sections 3b and 3c, we present results that aid interpretation of the discharge forecast skills and errors with focus on the potential multimodel improvements:

- the four uncorrected TIGGE forecasts with ERAI-Land initialization;
- the MM combination of the four uncorrected models with the ERAI, ERAI-Land, and MERRA-Land initializations and the grand ensemble of these three MM combinations (called GMM hereafter);
- the GMM combinations of the 30-day-corrected, the initial-time-corrected, and the combined initial-time- and 30-day-corrected MM forecasts (first initial-time-correct the forecasts, then apply the 30-day correction on these); and
- finally, the GMM of the BMA combined MM forecasts (from all three initializations) with the uncorrected models, the initial-time-corrected models, and also the uncorrected models extended by the persistence as a separate single value model.

FIG. 3. Impact of different forcing configurations in HTESSEL on the discharge outputs as a relative change compared to baseline. The black dashed line displays the impact of changing the LML to surface forcing (2 m for temperature and humidity, 10 m for wind). The colored lines highlight the impact of replacing different EC forcing variables, either individually or in combination, with ERAI-Land data.

a. Discharge sensitivity to meteorological forcing

The impact of replacing HTESSEL forcing variables other than precipitation (combination of Rad, Wind, and THP tests) with ERAI-Land (Fig. 3) is rather small ($\sim 3\%$ by $T + 240$, brown curve in Fig. 3). The least influential is the Wind (red curve), while the biggest contribution comes from the THP (green curve). When all ensemble forcing is replaced, including precipitation, the impact jumps to $\sim 15\%$ by $T + 240$ h, showing that a large majority of the change in the discharge comes from differences in precipitation (not shown).

The analysis of different areas and periods (see Table 2) highlights that larger impacts are seen for the winter period where the contribution of precipitation decreases and the contribution of the other forcing variables, both individually and combined, increases by approximately twofold to fivefold (this is particularly noticeable for THP). This is most likely a consequence of the snow-related processes, with snowmelt being dependent on temperature, radiation, and also wind in the cold seasons.

This also implies that the results are dependent on seasonality, a result that was also found by Liu et al. (2013), who looked at the skill of postprocessed precipitation forecasts using TIGGE data for the Huai River basin in China. In this study, because of the relatively short period we were able to use in the forecasts verification, scores were only computed for the whole verification period and no seasonal differences were analyzed.

Regarding the change from LML to surface forcing for temperature (2 m), wind (10 m), and humidity (2 m), the potential impact can be substantial, as shown by an example for 1–10 January 2012 in Fig. 4. In such cold winter conditions, large erroneous surface runoff values could appear in some parts of Russia when switching to surface forcing in HTESSEL. The representation of dew deposition is a general feature of HTESSEL that can be amplified in stand-alone mode. When coupled to the atmosphere, the deposition is limited in time, as it leads to a decrease of atmospheric humidity. However, in stand-alone mode, since the atmospheric conditions are prescribed, large deposition rates can be generated when the atmospheric forcing is not in balance (e.g., after model grid interpolation or changing from LML to surface forcing).

This demonstrates that with a land surface model such as HTESSEL, particular care needs to be taken in design of the experiments when model imbalances are expected. The use of surface data was an acceptable compromise as the sensitivity experiments highlighted only a small impact caused by the switch from LML to surface forcing (black dashed line in Fig. 3), and similarly by the impact of the Rad test, confirming that the necessary changes in the TIGGE land surface model setup did not have a major impact on the TIGGE discharge.

b. Reanalysis impact on discharge

The quality of the historical river flow that provides initial conditions for the CaMa-Flood TIGGE routing is expected to have a significant impact on the forecast skill. We analyze the discharge performance that is

TABLE 2. Detailed evaluation of the discharge sensitivity experiments at $T + 240$ h range for different areas and periods. Relative discharge differences are shown after replacing EC forcing variables, either individually or in combination, by ERAI-Land, and also the LML with surface forcing (2 m for temperature and humidity, 10 m for wind). The whole globe, the northern extratropics (defined here as 35° – 70° N), and the tropics (30° S– 30° N) as well as the specific seasons are displayed.

Avg diff (%)	Rad	THP	Wind	Rad + THP + Wind	Rad + THP + Wind + Prec	Surf vs LML
Global	1.0	2.4	0.6	2.9	15.6	0.7
Northern extratropics	1.0	3.1	0.6	3.5	12.8	0.8
Northern extratropics JJA	0.5	0.6	0.2	0.9	13.0	0.3
Northern extratropics DJF	1.1	2.9	0.7	3.6	9.5	0.8
Tropics	1.0	1.1	0.5	1.8	17.6	0.5
Tropics JJA	0.9	0.9	0.4	1.5	15.0	0.5
Tropics DJF	1.2	1.3	0.6	2.1	18.7	0.6

FIG. 4. Surface runoff output of HTESSEL for the period 1–10 Jan 2012 (240-h accumulation) from two EC experiments, using (a) surface forcing and (b) LML forcing, where possible. In (a), very large erroneous surface runoff values appear in very cold winter conditions.

highlighted in Fig. 5. This shows the MAESS and CORR for the ERAI-, ERAI-Land-, and MERRA-Land-simulated historical discharge from 1981 to 2010, and for their equal-weight multimodel average (MMA). The results are provided as continental and also as global averages of the available stations for Europe (~150 stations), North America (~350 stations), South America (~150 stations), Africa (~80 stations), Asia

(mainly Russia, 60 stations), and Australia and Indonesia (~50 stations), making ~840 stations globally.

The general quality of these global simulations is quite low. The MAESS averages over the available stations (see Fig. 1) are <0 for all continents, that is, large-scale average performance is worse than the daily observed climatology. The models are closest to the observed climate performance over Europe and Australia and

FIG. 5. Historical discharge forecast performance for ERAI-Land, ERAI, MERRA-Land, and their equal-weight MMA. MAESS and CORR are provided for each continent (North America, South America, Europe, Africa, Asia, and Australia and Indonesia). The reference forecast system in the skill score is the observed discharge climate as daily prediction. CORR are also provided for the observed climate. The scores are continental and global averages of the individual scores of the available stations (for station reference, see Fig. 1).

FIG. 6. Relative improvements in CORR by equal-weight average of ERAI-Land, ERAI, and MERRA-Land discharges. Values show the change in CORR compared with the best of ERAI-Land, ERAI, and MERRA-Land. Positive values show improvement while negative change means lower skill in the average than in the best of the three historical discharges.

Indonesia. The correlation between the simulated and observed time series shows a slightly more mixed picture, at least in some cases; especially Europe and Australia and Indonesia, the model is better than the observed climate. It is interesting to note that although the observed climate produces a high forecast time series correlation, in Asia the reanalysis discharge scores very low for all three sources. This could be related to the problematic handling of the snow in that area.

Figure 2 shows an example where MERRA-Land displayed a very strong negative bias. This example highlights the large variability among these data sources and is not an indication of the overall quality. Although MERRA-Land shows generally negative bias (not shown), the overall quality of the three reanalysis-driven historical discharge datasets is rather comparable. The highest skill and correlation is generally shown by ERAI-Land for most of the regions with the exception of Africa and Australia, where MERRA-Land is superior. ERAI, as the oldest dataset, appears to be the least skillful. Reichle et al. (2011) have found the same relationship between MERRA-Land and ERAI using 18 catchments in the United States. Although they computed correlation between seasonal anomaly time series (rather than the actual time series evaluated here), they could show that runoff estimates had higher correlation of the anomaly time series in MERRA-Land than in ERAI.

The multimodel average of the three simulations is clearly superior in the global and also in the continental

averages, with very few exceptions that have marginally lower MMA scores compared with the best individual reanalysis. The MMA is able to improve on the best of the three individual datasets at about half of the stations globally, both in the MAESS and CORR. Figure 6 shows the improvements in correlation. The points where the combination of the three reanalyses helps to improve on the best model cluster are mainly over Europe, Amazonia, and the eastern United States. On the other hand, the Northern Hemisphere winter areas seem to show mainly deterioration. This again is most likely related to the difficulty in the snow-related processes, which can hinder the success of the combination if, for example, one model is significantly worse with larger biases than the other two. Further analysis could help identify these more detailed error characteristics, providing a basis for further potential improvements.

c. Improving the forecast distribution

Figure 7 displays example hydrographs of some analyzed forecast products for a single forecast run to provide a practical impression of our experiments. The forecasts from 18 April 2009 are plotted for the GRDC station of Lobith in the Netherlands. The thin solid colored lines are the four TIGGE models (ECMWF, UKMO, NCEP, and CMA) plotted together (MM) with ERAI-Land (red), ERAI (green), and MERRA-Land (blue) initializations. They start from very different levels that are quite far from the observation (thick black line), but then seem to converge to roughly the

FIG. 7. Example of different discharge forecast products for the GRDC station of Lobith on the Rhine River in the Netherlands. All forecasts are from the run at 0000 UTC 18 Apr 2009 up to $T + 240$ h. The following products are plotted: multimodel combinations of four TIGGE models (ECMWF, UKMO, NCEP, and CMA) with ERAI-Land (solid red lines), ERAI (solid green lines), and MERRA-Land (solid blue lines) initializations; 30-day-corrected (dashed-dotted lines), initial-time-corrected (dashed lines), and 30-day- and initial-time-corrected (dotted lines) versions of the three multimodel combinations, each with all three initializations (with the respective colors); and finally, the BMA versions of the three multimodel combinations (all with gray lines, only from $T + 24$ h). The verifying observations are displayed by the black line.

same range in this example. The ensemble mean of the initial error-corrected MMs (from the three initializations with dashed lines), which by definition start from the observed discharge at $T + 0$ h, then follow faithfully the pattern of the mean of the respective MMs. The 30-day-corrected forecasts (dashed-dotted lines) follow a pattern relative to the MM ensemble means set by the performance of the last 30 days. The combination of the two bias-correction methods (dotted lines) blends the characteristics of the two; all three versions start from the observation (as first the initial error is removed) and then follow the pattern set by the past 30-day performance of this initial time-corrected forecast. Finally, the BMA-transformed (uncorrected) MMs (thin gray lines) happen to be closest to the observations in this example, showing a rather uniform spread throughout the processed range from $T + 24$ h to $T + 240$ h.

The quality of the TIGGE discharge forecasts based on the verified period from August 2008 to May 2010 is strongly dependent on the historical discharge that is used to initialize them. Figure 5 highlighted that the daily observed discharge climate is a better predictor than any of the three historical reanalysis-driven discharges ($\text{MAESS} < 0$). It is therefore not surprising that the uncorrected TIGGE forecasts show similarly low relative skill based on the CRPS (Fig. 8). Figure 8 also

FIG. 8. Discharge forecast performance for forecast ranges from $T + 0$ h to $T + 240$ h from August 2008 to May 2010 as global averages of CRPSS (computed at each station over the whole period) with the following forecast products. Gray lines indicate the four TIGGE models (ECMWF, UKMO, NCEP, and CMA) with ERAI-Land initialization, and a multimodel combination of these four models with ERAI-Land (red line), ERAI (green line), and MERRA-Land (blue line) initialization is also shown. The orange line represents a grand combination of these three multimodels, and grand combinations for six postprocessed products are shown: the multimodel of the 30-day correction (burgundy dashed line), the initial error correction (purple dashed line with markers), the 30-day and initial error correction combination (solid burgundy line with markers), two BMA versions of the multimodel—one with the uncorrected forecasts (black line without markers) and one with the uncorrected forecasts extended by the persistence as predictor (black line with circles)—and the persistence forecast (light blue dashed line with circles). The CRPSS is positively oriented and has a perfect value of 1. The 0 value line represents the quality of the reference system, the daily observed discharge climate.

shows the performance of the four models (gray dashed lines). In this study, we concentrate on the added value of the multimodel combination and do not distinguish between the four raw models. The scores change very little over the 10-day forecast period, showing a marginal increase in CRPSS as lead time increases. This is indicative of the incorrect initialization, with the forecast outputs becoming less dependent on initialization further into the medium range, and slowly converging toward climatology.

The first stage of the multimodel combination is the red line in Fig. 8, the combination of the uncorrected four models with the same ERAI-Land initialization. On the basis of this verification period and global station list, the simple equal-weight combination of the ensembles does not really seem to be able to improve on the best model. However, we have to acknowledge that the performance in general is very low.

The other area where we expect improvements through the multimodel approach is the initialization. Figure 8 highlights a significant improvement when using

three historical discharge initializations instead of only one. The quality of the ERAI-Land (red), ERAI (green), and MERRA-Land (blue) initialized forecasts (showed here only the multimodel combination versions) are comparable, with the ERAI-Land slightly ahead, which is in agreement with the results of the direct historical discharge comparisons presented in [section 4b](#). However, the grand combination of the three is able to improve significantly (orange line) on all of them. The improvement is much larger at shorter lead times as the TIGGE meteorological inputs provide lower spread, and therefore the spread introduced by the different initializations is able to have a bigger impact.

The quality of the discharge forecasts could be improved noticeably by introducing different initial conditions. However, the CRPSS is still significantly below 0, pointing to the need for postprocessing. In this study, we have experimented with a few methods that were proven to be beneficial.

The 30-day correction removed the mean bias of the most recent 30 runs from the forecasts. [Figure 8](#) shows the grand combination of the 30-day bias-corrected multimodels (with all three initializations), which brings the CRPSS to almost 0 throughout the 10-day forecast range (burgundy dashed line in [Fig. 8](#)). This confirms that the forecasts are severely biased. In addition, the shape of the curve remains fairly horizontal, suggesting this correction is not making the best use of the temporal patterns in the bias.

Further significant improvements in CRPSS are gained at shorter forecast ranges by using the initial time correction (purple dashed line with markers in [Fig. 8](#)), which does make use of temporal patterns in the bias. The shape of this error curve shows a typical pattern with the CRPSS decreasing with forecast range, reflecting the decreasing impact of the initial time correction and increased uncertainty in the forecast. The impact of the initial time errors gradually decreases until it finally disappears by around day 5 or 6, when the 30-day correction becomes superior.

The combination of the two methods, by applying the 30-day bias correction to forecasts already adjusted by the initial time correction, blends the advantages of both corrections. The CRPSS is further improved mainly in the middle of the 10-day forecast period with disappearing gain by $T + 240$ h (solid burgundy line with markers in [Fig. 8](#)).

The fact that the performance of the 30-day correction is worse in the short range than the initial time correction highlights that the impact of the errors at initial time has a structural component that cannot be explained by the temporally averaged bias. Similarly, the initial time correction cannot account exclusively

for the large biases in the forecasts as its impact trails off relatively quickly.

The persistence forecast shows a distinct advantage over these postprocessed forecasts (light blue dashed line with circles in [Fig. 8](#)). There is positive skill up to $T + 144$ h and the advantage of the persisted observation as a forecast diminishes, so that by $T + 240$ h its skill is similar to that of the combined corrected forecasts. This further highlights that the utilization of the discharge observations in the forecast production promises to provide a really significant improvement.

It is suggested that the structure of the initial errors has two main components: (i) biases in the reanalysis initializations due to biases in the forcing (e.g., precipitation) and in the simulations (e.g., evapotranspiration) and (ii) biases introduced by timing errors in the routing model due, in part, to the lack of optimized model parameters. A further evaluation of the weight of each of these error sources is beyond the scope of this study.

The final of our trialed postprocessing methods is the BMA. In [Fig. 8](#), similarly to the other postprocessed products, only the grand combination is displayed of the three BMA-transformed MMs with the different initializations. The BMA of the uncorrected forecasts was able to increase further the CRPSS markedly across all forecast ranges except $T + 24$ h (black line without markers). The results for $T + 24$ h suggest that at this lead time the perfect initial error correction from $T + 0$ h still holds superior.

The other two BMA versions, one with the uncorrected forecasts extended by the persistence as predictor (black line with circles) and one with the initial-time-corrected forecasts (not shown), both provide further skill improvements. The one with the persistence performs overall better, especially in the first few days. The BMA incorporating the persistence forecast remains skillful up to $T + 168$ h, the longest lead time of any of the forecast methods tested. At longer lead times (days 8–10) the BMA of the uncorrected model forecast appears to provide the highest skill of all the postprocessed products. This is evidence that the training of the BMA is not optimal. This is in part due to the estimation methodology used. More significantly experiments (not reported) show that the optimal training window for the BMA varies across sites, showing a different picture for the BMA with or without persistence, and also delivering potentially higher global average skill using a longer window.

Although [Fig. 8](#) shows only the impact of the four postprocessing methods on the grand combination of the MM forecasts, the individual MMs with the three initializations show the same behavior. The GMMs

FIG. 9. Distribution of the skill increments over all stations provided by six combination and postprocessing products for two time ranges: (a) $T + 24$ h and (b) $T + 240$ h. The x axis shows the reference skill, the average CRPSS of the four TIGGE models with the ERA-Land initialization, while the y axis displays the CRPSS of the postprocessed forecasts at the stations. The six products are the MM combination of the four models with ERAI-Land (red circles), the GMM combination of the three uncorrected MMs (green triangles), and the GMM combination of four postprocessed products: the 30-day-corrected MMs (orange triangles), the initial-time-corrected MMs (cyan squares), the combined 30-day- and initial-time-corrected MMs (purple stars), and finally the BMA-transformed MMs of the uncorrected forecasts (blue stars), where all the MMs are the three MM with the different initializations. The diagonal line represents no skill improvement; above this line the six products are better, while below it they are worse than the reference. The CRPSS values are computed based on the period from August 2008 to May 2010. Some of the stations that have reference CRPSS below -5 are not plotted.

always outperform the three MMs for all the postprocessing products; for example, for the most skillful method, the BMA, the grand combination extends the positive skill by ~ 1 day (from around 5 days to 6 days, not shown).

The distribution of the skill increments over all stations provided by different combination and postprocessing products is summarized in Fig. 9 at $T + 24$ h (Fig. 9a) and $T + 240$ h (Fig. 9b). The reference skill is the average CRPSS of the four TIGGE models with the ERA-Land initialization (these values are represented by the gray dashed lines in Fig. 8). Figure 9 highlights the structure of the improvements in different ranges of the CRPSS for the different methods over all verified stations in the period from August 2008 to May 2010. The picture is characteristically different at different lead times, as suggested by the $T + 24$ h and $T + 240$ h plots. At short range, the improvements of the different products scale nicely into separate bands. The relatively simple MM combination of the four models with ERAI-Land (red circles) does not improve on the forecast; the increments are small and with mixed sign. The GMM combination of the three uncorrected MMs (green triangles) shows a marked improvement, and the 30-day correction version (orange triangles) improves further while the initial time correction products (cyan squares and purple stars) show the largest improvement over most of the stations. At this short $T + 24$ h range, the

BMA (blue stars) of the uncorrected forecasts is slightly behind, which is a general feature across the displayed CRPSS range from -5 to 1.

In contrast to the short range, $T + 240$ h provides a significantly different picture. The relatively clear ranking of the products is gone by this lead time. The MM and GMM combinations are able to improve slightly for most of the stations, but at this range the contribution seems to be generally always positive. The postprocessing methods at this medium range, however, deteriorate the forecasts sometimes, especially in the range from -1 to 0.5 (the 30-day correction seems to behave noticeably better in this respect). The general improvements are clear though for most of the stations, and also the overall ranking of the methods seen in Fig. 8 is reflected, although much less clearly than at $T + 24$ h, with the BMA topping the list at $T + 240$ h.

Finally, Fig. 10 presents the discharge performance we could achieve in this study for all the stations that could be processed in the period from August 2008 to May 2010 at $T + 240$ h. It displays the CRPSS of the best overall product, the GMM with the BMA of the uncorrected forecasts (combination of the three BMA-transformed MMs with the three initializations without initial time or 30-day bias correction). The variability of the scores is very large geographically, but there are emerging patterns. Higher performance is observed in the Amazon and in central and western parts of the

FIG. 10. Global CRPSS distribution of the highest quality postprocessed product at $T + 240$ h, the grand multimodel combination of the BMA-transformed uncorrected forecasts, based on the period from August 2008 to May 2010. The CRPSS is positively oriented and has a perfect value of 1. The 0 value represents the quality of the reference system, the daily observed discharge climate.

United States, while lower CRPSSs are seen over the Rocky Mountains in North America and in northerly points in Europe and Russia. Unfortunately, the geographical coverage of the stations is not good enough to draw more detailed conclusions.

5. Conclusions

This study has shown aspects of building a global multimodel hydrometeorological forecasting system using the TIGGE archive and analyzed the impact of the postprocessing required to run a multimodel system on the forecasts.

The atmospheric input was taken from four operational global meteorological ensemble systems, using data available from TIGGE. The hydrological component of this study was the HTESSEL land surface model while the CaMa-Flood global river-routing model was used to integrate runoff over the river network. Observations from the GRDC discharge archive were used for evaluation and postprocessing.

We have shown that the TIGGE archive is a valuable resource for river discharge forecasting, and three main objectives were successfully addressed: (i) the sensitivity of the forecasting system to the meteorological input variables, (ii) the potential improvements to the historical discharge dataset (which provides initial river conditions to the forecast routing), and (iii) improving the predictive distribution of the forecasts. The main outcomes can be grouped as follows:

- (i) The impact of replacing or altering the input meteorological variables to fit the system requirements is small and allows the use of variables from the TIGGE archive for this hydrological study.
- (ii) The multimodel average historical discharge dataset provides a very valuable source of uncertainty and a general gain in skill.
- (iii) Significant improvements in the forecast distribution can be produced through the use of initial time and 30-day bias corrections on the TIGGE model discharge, or on the combination of the forecast models; however, the combination of techniques used has a big impact on the improvement observed, with the best BMA products providing positive skill up to 6 days.

The quality of the raw TIGGE-based discharge forecasts has been shown to be low, mainly determined by the limited performance of the reanalysis-driven historical river conditions analyzed in [section 4b](#). The lower skill is in agreement with results found in other studies. For example, [Alfieri et al. \(2013\)](#) showed that in the context of GloFAS, the LISFLOOD hydrological model ([Van Der Knijff et al. 2010](#)), forced by ERAI-Land runoff, shows variable performance based on the 1990–2010 historical period. From the analyzed 620 global observing stations the Pearson's correlation coefficient reaches as low as -0.2 , and only 71% of them provide correlation values above 0.5. [Donnelly et al. \(2015\)](#) highlighted similar behavior with the E-HYPE system based on 181 river gauges in Europe for 1981–2000. The

correlation component of the Kling–Gupta efficiency started around 0, and geographical distribution of values in Europe was very similar to our result (not shown). The lowest correlation was found mainly in Spain and in Scandinavia, with a comparable average value to our European mean of 0.6–0.7 (see Fig. 5).

The combination and postprocessing methods we applied to the discharge forecasts provided significant improvement of the skill. Although the simple multimodel combinations and the 30-day bias correction (removing the mean error of the most recent 30 days) both provide significant improvements, they are not capable of achieving positive global skill (i.e., outperform the daily observed discharge climate). The initial time correction, by adjusting to the observations at initial time and applying this error correction into the forecast, is able to provide skill in the short range (only up to 2–3 days), especially when combined with the 30-day correction. However, the impact quickly wears off and for longer lead times (up to about 6 days) only the BMA postprocessing method is able to provide positive average global skill (closely followed by the persistence).

Although other studies could show significant improvement by using multiple meteorological inputs (e.g., Pappenberger et al. 2008), in this study the impact of combining different TIGGE models is rather small. This is most likely a consequence of the overwhelming influence of the historical river conditions on the river initialization. The grand combinations, when we combine the forecasts produced with different reanalysis-driven historical river conditions, however, always outperform the individual MMs (single initialization) for all the postprocessing products. They provide a noticeable overall skill improvement, which in our study translated into an extension of the lead time, when the CRPSS drops below 0, by about one day as a global average for the most skillful BMA forecasts.

In the future we plan to extend this study to address other aspects of building a skillful multimodel hydrometeorological system. The following areas are considered:

- (i) Include other datasets that provide global coverage of runoff data on high enough horizontal resolution, such as the Japanese 55-year Reanalysis (JRA-55; Kobayashi et al. 2015) or the NCEP Climate Forecast System Reanalysis (CFSR; Saha et al. 2010) to provide further improvements in the initial river condition estimates.
- (ii) Introduce the multihydrology aspect by adding an additional land surface model such as the Joint UK Land Environment Simulator (JULES; Best et al. 2011).

- (iii) The presented scores in this study are relatively low even with the postprocessing methods applied. To achieve significantly higher overall scores, the information on the discharge observations should be utilized in the modeling.
- (iv) Similarly, the discharge quality could be significantly improved by better calibration of many of the watersheds in the CaMa-Flood routing.
- (v) Alternatively, the application of different river-routing schemes such as LISFLOOD, which is currently used in the GloFAS, would also provide potential increase in the skill through the multimodel use.
- (vi) Further analysis of the errors and the trialing of other postprocessing methods could also lead to potential improvements. In particular, better allowance should be made for temporal correlation in the forecast errors. The use of the extreme forecast index (Zsótér 2006) as a tool to compare the forecasts to the model climate could potentially bring added skill into the flood predictions.

Acknowledgments. We are thankful to the European Commission for funding this study through the Global Earth Observation System of Systems (GEOSS) Interoperability for Weather, Ocean, and Water (GEOWOW) project in the 7th Framework Programme for Research and Technological Development (FP7/2007–2013) under Grant Agreement 282915. We are also grateful to the Global Runoff Data Centre in Koblenz, Germany, for providing the discharge observation dataset for our discharge forecast analysis.

REFERENCES

- Agustí-Panareda, A., G. Balsamo, and A. Beljaars, 2010: Impact of improved soil moisture on the ECMWF precipitation forecast in West Africa. *Geophys. Res. Lett.*, **37**, L20808, doi:10.1029/2010GL044748.
- Ajami, N. K., Q. Duan, and S. Sorooshian, 2007: An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resour. Res.*, **43**, W01403, doi:10.1029/2005WR004745.
- Alfieri, L., P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, and F. Pappenberger, 2013: GloFAS—Global ensemble streamflow forecasting and flood early warning. *Hydrol. Earth Syst. Sci.*, **17**, 1161–1175, doi:10.5194/hess-17-1161-2013.
- Balsamo, G., A. Beljaars, K. Scipal, P. Viterbo, B. van den Hurk, M. Hirschi, and A. K. Betts, 2009: A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the Integrated Forecast System. *J. Hydrometeorol.*, **10**, 623–643, doi:10.1175/2008JHM1068.1.
- , S. Boussetta, P. Lopez, and L. Ferranti, 2010: Evaluation of ERA-Interim and ERA-Interim-GPCP-rescaled precipitation

- over the U.S.A. ERA Rep. 01/2010, ECMWF, 10 pp. [Available online at <http://www.ecmwf.int/sites/default/files/elibrary/2010/7926-evaluation-era-interim-and-era-interim-gpcp-rescaled-precipitation-over-usa.pdf>.]
- , F. Pappenberger, E. Dutra, P. Viterbo, and B. van den Hurk, 2011: A revised land hydrology in the ECMWF model: A step towards daily water flux prediction in a fully-closed water cycle. *Hydrol. Processes*, **25**, 1046–1054, doi:[10.1002/hyp.7808](https://doi.org/10.1002/hyp.7808).
- , and Coauthors, 2015: ERA-Interim/Land: A global land surface reanalysis data set. *Hydrol. Earth Syst. Sci.*, **19**, 389–407, doi:[10.5194/hess-19-389-2015](https://doi.org/10.5194/hess-19-389-2015).
- Bao, H. H., and O. Zhao, 2012: Development and application of an atmospheric–hydrologic–hydraulic flood forecasting model driven by TIGGE ensemble forecasts. *Acta Meteor. Sin.*, **26**, 93–102, doi:[10.1007/s13351-012-0109-0](https://doi.org/10.1007/s13351-012-0109-0).
- Best, M. J., and Coauthors, 2011: The Joint UK Land Environment Simulator (JULES), model description—Part 1: Energy and water fluxes. *Geosci. Model Dev.*, **4**, 677–699, doi:[10.5194/gmd-4-677-2011](https://doi.org/10.5194/gmd-4-677-2011).
- Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, doi:[10.1175/2010BAMS2853.1](https://doi.org/10.1175/2010BAMS2853.1).
- Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.*, **131**, 2131–2150, doi:[10.1256/qj.04.71](https://doi.org/10.1256/qj.04.71).
- Cane, D., S. Ghigo, D. Rabuffetti, and M. Milelli, 2013: Real-time flood forecasting coupling different postprocessing techniques of precipitation forecast ensembles with a distributed hydrological model. The case study of May 2008 flood in western Piemonte, Italy. *Nat. Hazards Earth Syst. Sci.*, **13**, 211–220, doi:[10.5194/nhess-13-211-2013](https://doi.org/10.5194/nhess-13-211-2013).
- Cloke, H. L., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *J. Hydrol.*, **375**, 613–626, doi:[10.1016/j.jhydrol.2009.06.005](https://doi.org/10.1016/j.jhydrol.2009.06.005).
- Coccia, G., and E. Todini, 2011: Recent developments in predictive uncertainty assessment based on the model conditional processor approach. *Hydrol. Earth Syst. Sci.*, **15**, 3253–3274, doi:[10.5194/hess-15-3253-2011](https://doi.org/10.5194/hess-15-3253-2011).
- Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, doi:[10.1002/qj.828](https://doi.org/10.1002/qj.828).
- Demargne, J., and Coauthors, 2014: The science of NOAA's Operational Hydrologic Ensemble Forecast Service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, doi:[10.1175/BAMS-D-12-00081.1](https://doi.org/10.1175/BAMS-D-12-00081.1).
- Dong, L., L. Xiong, and K. Yu, 2013: Uncertainty analysis of multiple hydrologic models using the Bayesian model averaging method. *J. Appl. Math.*, **2013**, 346045, doi:[10.1155/2013/346045](https://doi.org/10.1155/2013/346045).
- Donnelly, C., J. C. M. Andersson, and B. Arheimer, 2015: Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. *Hydrol. Sci. J.*, **61**, 255–273, doi:[10.1080/02626667.2015.1027710](https://doi.org/10.1080/02626667.2015.1027710).
- Dutra, E., C. Schär, P. Viterbo, and P. M. A. Miranda, 2011: Land-atmosphere coupling associated with snow cover. *Geophys. Res. Lett.*, **38**, L15707, doi:[10.1029/2011GL048435](https://doi.org/10.1029/2011GL048435).
- Emerton, R. E., and Coauthors, 2016: Continental and global scale flood forecasting systems. *Wiley Interdiscip. Rev.: Water*, **3**, 391–418, doi:[10.1002/wat2.1137](https://doi.org/10.1002/wat2.1137).
- Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138**, 190–202, doi:[10.1175/2009MWR3046.1](https://doi.org/10.1175/2009MWR3046.1).
- Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151, doi:[10.1146/annurev-statistics-062713-085831](https://doi.org/10.1146/annurev-statistics-062713-085831).
- Haddeland, I., and Coauthors, 2011: Multimodel estimate of the global terrestrial water balance: Setup and first results. *J. Hydrometeor.*, **12**, 869–884, doi:[10.1175/2011JHM1324.1](https://doi.org/10.1175/2011JHM1324.1).
- Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1814–1827, doi:[10.1002/qj.1895](https://doi.org/10.1002/qj.1895).
- He, Y., F. Wetterhall, H. L. Cloke, F. Pappenberger, M. Wilson, J. Freer, and G. McGregor, 2009: Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. *Meteor. Appl.*, **16**, 91–101, doi:[10.1002/met.132](https://doi.org/10.1002/met.132).
- , and Coauthors, 2010: Ensemble forecasting using TIGGE for the July–September 2008 floods in the Upper Huai catchment: A case study. *Atmos. Sci. Lett.*, **11**, 132–138, doi:[10.1002/asl.270](https://doi.org/10.1002/asl.270).
- Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, doi:[10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2).
- Huffman, G. J., R. F. Adler, D. T. Bolvin, and G. Gu, 2009: Improving the global precipitation record: GPCP version 2.1. *Geophys. Res. Lett.*, **36**, L17808, doi:[10.1029/2009GL040000](https://doi.org/10.1029/2009GL040000).
- Kobayashi, S., and Coauthors, 2015: The JRA-55 Reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48, doi:[10.2151/jmsj.2015-001](https://doi.org/10.2151/jmsj.2015-001).
- Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, doi:[10.1126/science.285.5433.1548](https://doi.org/10.1126/science.285.5433.1548).
- Krzysztofowicz, R., 1997: Transformation and normalization of variates with specified distributions. *J. Hydrol.*, **197**, 286–292, doi:[10.1016/S0022-1694\(96\)03276-3](https://doi.org/10.1016/S0022-1694(96)03276-3).
- Liang, Z., D. Wang, Y. Guo, Y. Zhang, and R. Dai, 2013: Application of Bayesian model averaging approach to multimodel ensemble hydrologic forecasting. *J. Hydrol. Eng.*, **18**, 1426–1436, doi:[10.1061/\(ASCE\)HE.1943-5584.0000493](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000493).
- Liu, Y., Q. Duan, L. Zhao, A. Ye, Y. Tao, C. Miao, X. Mu, and J. C. Schaake, 2013: Evaluating the predictive skill of post-processed NCEP GFS ensemble precipitation forecasts in China's Huai River basin. *Hydrol. Processes*, **27**, 57–74, doi:[10.1002/hyp.9496](https://doi.org/10.1002/hyp.9496).
- Olsson, J., and G. Lindström, 2008: Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. *J. Hydrol.*, **350**, 14–24, doi:[10.1016/j.jhydrol.2007.11.010](https://doi.org/10.1016/j.jhydrol.2007.11.010).
- Pappenberger, F., J. Bartholmes, J. Thielen, H. L. Cloke, R. Buizza, and A. de Roo, 2008: New dimensions in early flood warning across the globe using grand-ensemble weather predictions. *Geophys. Res. Lett.*, **35**, L10404, doi:[10.1029/2008GL033837](https://doi.org/10.1029/2008GL033837).
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:[10.1175/MWR2906.1](https://doi.org/10.1175/MWR2906.1).
- Reichle, R. H., R. D. Koster, G. J. M. De Lannoy, B. A. Forman, Q. Liu, S. P. P. Mahanama, and A. Toure, 2011: Assessment and enhancement of MERRA land surface hydrology estimates. *J. Climate*, **24**, 6322–6338, doi:[10.1175/JCLI-D-10-05033.1](https://doi.org/10.1175/JCLI-D-10-05033.1).

- Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1057, doi:[10.1175/2010BAMS3001.1](https://doi.org/10.1175/2010BAMS3001.1).
- Schefzik, R., L. T. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, doi:[10.1214/13-STS443](https://doi.org/10.1214/13-STS443).
- Thielen, J., J. Bartholmes, M. H. Ramos, and A. de Roo, 2009: The European Flood Alert System—Part 1: Concept and development. *Hydrol. Earth Syst. Sci.*, **13**, 125–140, doi:[10.5194/hess-13-125-2009](https://doi.org/10.5194/hess-13-125-2009).
- Todini, E., 2008: A model conditional processor to assess predictive uncertainty in flood forecasting. *Int. J. River Basin Manage.*, **6**, 123–137, doi:[10.1080/15715124.2008.9635342](https://doi.org/10.1080/15715124.2008.9635342).
- Van Der Knijff, J. M., J. M. J. Younis, and A. P. J. De Roo, 2010: LISFLOOD: A GIS-based distributed model for river basin scale water balance and flood simulation. *Int. J. Geogr. Inf. Sci.*, **24**, 189–212, doi:[10.1080/13658810802549154](https://doi.org/10.1080/13658810802549154).
- Velázquez, J. A., F. Anctil, M. H. Ramos, and C. Perrin, 2011: Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. *Adv. Geosci.*, **29**, 33–42, doi:[10.5194/adgeo-29-33-2011](https://doi.org/10.5194/adgeo-29-33-2011).
- Vrugt, J. A., and B. A. Robinson, 2007: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resour. Res.*, **43**, W01411, doi:[10.1029/2005WR004838](https://doi.org/10.1029/2005WR004838).
- Yamazaki, D., S. Kanae, H. Kim, and T. Oki, 2011: A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resour. Res.*, **47**, W04501, doi:[10.1029/2010WR009726](https://doi.org/10.1029/2010WR009726).
- Zsótér, E., 2006: Recent developments in extreme weather forecasting. *ECMWF Newsletter*, ECMWF, Reading, United Kingdom, 8–17. [Available online at <http://www.ecmwf.int/sites/default/files/elibrary/2006/14618-newsletter-no107-spring-2006.pdf>.]