Inter-comparison between retrospective ensemble streamflow forecasts using meteorological inputs from ECMWF and NOAA/ESRL in the Hudson River sub-basins during Hurricane Irene (2011)
F. Saleh, V. Ramaswamy, N. Georgas, A. F. Blumberg and J. Pullen

ABSTRACT

The objective of this work was to evaluate the benefits of using multi-model meteorological ensembles in representing the uncertainty of hydrologic forecasts. An inter-comparison experiment was performed using meteorological inputs from different models corresponding to Hurricane Irene (2011), over three sub-basins of the Hudson River basin. The ensemble-based precipitation inputs were used as forcing in a hydrological model to retrospectively forecast hourly streamflow, with a 96-hour lead time. The inputs consisted of 73 ensemble members, namely one high-resolution ECMWF deterministic member, 51 ECMWF members and 21 NOAA/ESRL (GEFS Reforecasts v2) members. The precipitation inputs were resampled to a common grid using the bilinear resampling method that was selected upon analysing different resampling methods. The results show the advantages of forcing hydrologic forecasting systems with multi-model ensemble forecasts over using deterministic and single model ensemble forecasts. The work showed that using the median of all 73 ensemble streamflow forecasts relatively improved the Nash-Sutcliffe Efficiency and lowered the biases across the examined sub-basins, compared with using the ensemble median from an individual model. This research contributes to the growing literature that demonstrates the promising capabilities of multi-model systems to better describe the uncertainty in streamflow predictions.

Key words | ECMWF, HEC-HMS, Hurricane Irene, NOAA/ESRL GEFS, spatial resolution, streamflow ensemble forecasts

INTRODUCTION

The increase in frequency and intensity of riverine flooding has led to a rise in the awareness of flood preparedness among decision makers, highlighting the need to improve and communicate flood forecasts effectively (Berkom et al. 2007; Demeritt et al. 2007; Pitt 2008; Weaver et al. 2014; Hoss & Fischbeck 2016). In this context, early warning systems are increasingly used to produce flood forecasts at time horizons varying from hours up to 15 days (e.g. de Roo et al. 2005; Rabuffetti & Barbero 2005; Thielen et al. 2009; Thiemig et al. 2010; Schellekens et al. 2011; Alfieri et al. 2013; Gochis et al. 2013; Adams 2015; Salas et al. 2018). Such hydrological flood forecasting systems ingest meteorological inputs obtained from Numerical Weather Prediction (NWP) models (Cloke & Pappenberger 2009). Meteorological inputs from NWP models are known to have uncertainties arising from parameterizations of physical processes in atmospheric models’ resolution and initial conditions (Krzysztofowicz 2001a; Bartholmes & Todini 2005; Cuo et al. 2011). These uncertainties cascade from NWP models through the hydrologic models to the flood forecasts (Pappenberger et al. 2005; Saleh et al. 2017). To apply a measure of uncertainty arising from the initial state and parameterization of the NWP...
models (Buizza 2008; Velázquez et al. 2011), the concept of using every member of the ensemble is advantageous, and recent studies show the benefits over deterministic meteorological forecasts (Hsiao et al. 2013; Mittermaier & Csima 2017). In the context of hydrological forecasting systems, an ensemble of NWP models is used to account for the effects of uncertainty arising from meteorological predictions (Krzysztofowicz 2001b; Bowler et al. 2008; Hamill et al. 2008; Thielen et al. 2008; Cloke & Pappenberger 2009; Pagano et al. 2013; Emerton et al. 2016). Advances in computational power, technology and scientific research have allowed significant improvements in operational streamflow forecasting using ensembles (Demeritt et al. 2007; Tang et al. 2013; Alfieri et al. 2014). Numerous forecasting agencies around the world are using NWP ensemble inputs of varying spatial and temporal resolutions identified by the areas of interest and data availability. For example, meteorological data from the European Centre for Medium-Range Weather Forecasts (ECMWF) are widely used as inputs to hydrological models to issue flood forecasts in different regions of the world (Roulin & Vannitsem 2005; Marsalek et al. 2006; Johnell et al. 2007; Roulin 2007; Vehviläinen 2007; He et al. 2010; Laugesen et al. 2011; Saleh et al. 2016). ECMWF meteorological inputs are also used in operational flood forecasting systems, to predict floods in ungaged basins in Africa (Thiemig et al. 2010, 2015). Rabuffetti & Barbero (2005) developed an operational forecasting framework in the Piemonte region of Italy using precipitation and temperature inputs from the Royal Meteorological Service. Hopson & Webster (2010) developed an operational forecasting system for the Ganges and Brahmaputra River basins in Bangladesh using a hydrologic model initialized with National Aeronautics and Space Administration (NASA) and National Oceanic and Atmospheric Administration (NOAA) precipitation inputs, and forecasts were produced using ECMWF ensemble forecast products.

In the United States, streamflow forecasts are provided by the NOAA Advanced Hydrologic Prediction Service (AHPS) using inputs from the National Weather Service (NWS) (McEnery et al. 2005). As an enhancement to the NWS’s baseline river forecasting system the Community Hydrologic Prediction System (CHPS), an experimental end-to-end Hydrologic Ensemble Forecast Service (HEFS) is used to address a wide range of water information and services such as flood risk and water resources management (Demargne et al. 2014). The utility of the HEFS was demonstrated using hindcasting experiments in four River Forecast Centres (RFCs) (Brown et al. 2014a, 2014b). Furthermore, the National Water Model (NWM) has recently become available to provide streamflow forecasts in 2.7 million reaches across the continental United States and overcome the disparity of the availability of forecasts at discrete locations (Maidment 2017; Salas et al. 2018). The core of the NWM is the National Centre for Atmospheric Research (NCAR)-supported community Weather Research and Forecasting Hydrologic model (WRF-Hydro), which ingests forcing from NWP forecast data.

In contrast to Ensemble Protection System (EPS) forecasts from a single forecast centre, which only address the uncertainties in initial conditions and stochastic physics (Roulin 2007), multi-model forecasts from multiple forecast centres can improve the simulation of uncertainties arising from numerical implementation and/or data assimilation (Pappenberger et al. 2008). In this context, studies highlight the benefits of using multiple models for flood forecasting, increasing the lead times, facilitating improved early warnings and decision making, and producing reliable results compared with a single EPS (Pappenberger et al. 2008; Zsótér et al. 2016) in addition to increasing the skill of the forecast (He et al. 2009; Bao & Zhao 2012). The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) project was initiated to enable advanced research and demonstration of the multi-model ensemble concepts and operationally implement such systems mainly for hydrologic applications (Bougeault et al. 2010). As an example of a multi-model forecasting system in the United States, the NWS Meteorological Model Ensemble Forecast System (MMEFS) uses ensemble forecasts from the operational Short Range Ensemble Forecasting (SREF), Global Ensemble Forecast System (GEFS), as well as North American Ensemble Forecast System (NAEFS) ensembles (Adams 2015). NAEFS ensembles include both National Center for Environmental Prediction (NCEP) ensemble forecasts and Canadian global forecast model ensembles with a consistent bias correction of the ensembles (Toth et al. 2005).

The objective of this work was to investigate the benefits of forcing a hydrologic prediction framework with
meteorological ensembles from multi-model NWPs. The investigation used ensemble-based precipitation and streamflow forecasts from ECMWF and NOAA/ESRL (GEFS Reforecast v2) for an extreme rainfall event, Hurricane Irene, which caused widespread damage primarily due to inland flooding. The meteorological inputs were resampled to common spatial resolution and used as forcing to a semi-distributed hydrologic model to aid the inter-comparison efforts. The uncertainty in the precipitation inputs and the ensemble-based streamflow reforecasts was visually and statistically quantified for three different sub-basins on the lower Hudson River Basin.

The first part of this paper describes the context and the selected sub-basins, followed by a summary of the meteorological datasets. Then, a general overview of the hydrologic model and evaluation criteria of the precipitation inputs and streamflow outputs is outlined. Finally, an inter-comparison is performed to quantify the uncertainty from the propagation of precipitation inputs into the hydrological streamflow forecasts during Hurricane Irene.

**MATERIALS AND METHODS**

**Study areas and context**

Three sub-basins of the Hudson River basin were selected based on their drainage areas and their respective catchment properties (Figure 1). The three sub-basins (Figure 1) represent the Saddle River (140 km²), Croton River (980 km²) and Wallkill River (1,800 km²). The Croton River is located in the northern part of the Hudson River basin while Saddle River lies in the southern Hudson region (Figure 1). The areas under study have varying degrees of land cover and range from highly urbanized to rural (Figure 1). For example, floodplains within the Saddle River are highly developed and contain a mix of residential and commercial structures. Anthropogenic impacts such as the diurnal fluctuations at low flow from sewage effluents entering the Saddle River upstream and pumping are not quantified in this study. Floodplains within the Croton River show intermediate development with control structures such as the New Croton reservoir, which supplies water to New York City (Burns et al. 2003). In contrast, floodplains along Wallkill River are less urbanized and mainly composed of farmlands. The flow at Wallkill River is also regulated by dams and diversions for municipalities and irrigation.

The study is focused on the 2011 Hurricane Irene, which caused substantial flood damage due to its unique meteorological characteristics coupled with the topographical and geological character of the region, in addition to the heavily saturated soil from the wet summer (Coch 2012). The 2011 hurricane season had the third highest number of named storms since record keeping, and although Irene had decreased to tropical storm intensity when it made its landfall, it did damage equal to the Hurricane of 1821 (Category 1) and Hurricane of 1893 (Category 2) hurricanes as well as Tropical Storm Floyd (Coch 2012). The accumulated precipitation from Hurricane Irene during 27–30 August 2011 was more than 300 mm in certain areas of the Hudson River basin and caused widespread inundations of streams throughout New York and New Jersey, resulting in peak stream flows exceeding the 100-year recurrence interval at many observed stream gauges and causing heavy property and road damage (Saleh et al. 2016). Within the study area, the floodplains of Saddle River experienced severe flooding numerous times, most notably during tropical storm Floyd in 1999 and Hurricane Irene (Hoppe & Watson 2012). Businesses, property and agriculture along Wallkill River were impacted due to Hurricane Irene and suffered losses as crops exposed to flood waters could not be sold due to the release of toxic material from the floodwaters (NYSERDA 2014).

**HEC-HMS hydrological model**

HEC-HMS, developed by the US Army Corps of Engineers (USACE 2015), is a conceptual semi-distributed hydrological model used in rainfall-runoff modelling and other hydrologic studies (Anderson et al. 2002; Neary et al. 2004; Knebl et al. 2005; Amenguai et al. 2009; Chu & Steinman 2009; Halwatura & Najim 2013; Meenu et al. 2013; Seyoum et al. 2013; Yang & Yang 2014; Akter & Ahmed 2015; Saleh et al. 2016). The structure of the model is defined using empirically derived parameters, which are adjustable, and these parameters exist for runoff, base flow and river routing (Feldman 2000). In this work, the spatial variability and characteristics of the basin were accounted for using the
Figure 1 | Locations of the considered sub-basins in the Hudson River basin along with the topography.
modified Clark (ModClark) distributed method (Kull & Feldman 1998).

Infiltration capacity in the model was quantified using the gridded curve number (CN) methodology derived by the Soil Conservation Service (SCS) (USDA 1986; Mishra & Singh 2013). The SCS CN model estimates precipitation excess as a function of cumulative precipitation, soil cover, land use and antecedent moisture content (Feldman 2000). Initial flow and the recession constants were used to account for groundwater contributions to streamflow (Chow 1959; Maidment 1992; Feldman 2000). The river flow routing was based on the Muskingum equations of the HEC-HMS model (USACE 2013).

Hydrological model datasets

The Hudson River basin hydrological model (Saleh et al. 2016) is part of a fully automated operational framework, producing four forecast cycles of ensemble river discharge per day at an hourly time step, and is used by the New York Harbor Observing and Prediction System (NYHOPS) (Bruno et al. 2006; Georgas et al. 2014). The model was initialized using observed river discharge from United States Geological Survey (USGS) gauging stations through the National Water Information System (NWIS).

The model used in this study was a subset of the regional scale Hudson River basin model implemented by Saleh et al. (2016). This model was delineated using HEC-GeoHMS (Fleming & Doan 2015) into sub-basins based on flow direction and accumulation using topography obtained from the USGS National Elevation Dataset (NED) (Gesch et al. 2002; Fleming & Doan 2013), land surface cover obtained from the US Department of Agriculture National Resource Conservation Service (NRCS) and soil data from the State Soil Geographic Database (STATSGO) (Miller & White 1998). Each sub-basin was then sub-divided into hydrologic response units with a specific gridded SCS CN representing the runoff response rate based on intersecting land use data-sets obtained from the USGS National Land Cover Dataset (NLCD) (Homer et al. 2012) and land cover with the soil data using the CN-grid tool in HEC-GeoHMS (Gassman et al. 2007). The time of concentration (2–8 hours) and percentage imperviousness at each sub-basin were derived from a Geographical Information System (GIS) dataset using HEC-GeoHMS 10.2. Observed river discharge from USGS gauging stations located at the outlets of Saddle River at Lodi, New Jersey (USGS ID 01391500), Croton River near New Croton Dam at Croton, New York (USGS ID 01357500) and Wallkill River at Gardiner, New York (USGS ID 01371500) were automatically obtained through the NWIS using the R dataRetrieval package (Hirsch & De Cicco 2013). These data were recorded at 15 minute intervals and imported to USACE Data Storage System Visual Utility Engine (HEC-DSSVue) (HEC 2009). To derive the baseflow recession constants for each sub-basin, 25 years of observed historical flow data were separated using an automated baseflow separation technique based on the R low flow statistics package ‘lfstat’ (Koffler & Laaha 2012). The derived recession constant values ranged from 0.55 to 0.95 for the three basins selected for this study. Given the application of the model to forecast short-term extreme events within a 96-h forecast horizon, observed conditions from the gauge stations were used to obtain the baseflow and initialize the HEC-HMS model.

Meteorological datasets and resampling methods

The North American Regional Reanalysis (NARR) (Mesinger et al. 2006) precipitation dataset was considered as the baseline for the inter-comparison. The retrospective streamflow forecast of Hurricane Irene used meteorological inputs from 21 GEFS ensemble members (Hamill et al. 2015; Zhou et al. 2017), 51 ensemble members from ECMWF and one high-resolution deterministic member (HRES) from the ECMWF-HRES model (Molteni et al. 1996). The NWP models differ from each other in terms of spatial resolutions and required resampling to a common spatial grid for use in the hydrologic model.

North American Regional Reanalysis (NARR)

NARR is an atmospheric and land surface dataset for the North American domain in which high-quality and detailed precipitation observations are assimilated into the atmospheric analysis to create a long-term, dynamically consistent, high-resolution climate dataset for the North American domain. It includes a 25-year retrospective period from 1979 to 2003 and is augmented by the construction and daily execution of a system for near-real-time
continuation of the NARR, known as the Regional Climate Data Assimilation System (R-CDAS). However, given the proximity of the study domain to Canada, it is important to address the fact that some discontinuities in precipitation exist at the US–Canada border after 2003. This can be attributed to the spatial density of the assimilated rain gauges in the two national observational datasets. The temporal resolution of the NARR data is 3 hours and the spatial resolution is 32 km (Mesinger et al. 2006). The NARR precipitation data are publicly available from the National Climatic Data Centre (NCDC) and are used in several hydrological studies. For instance, Choi et al. (2009) used the NARR dataset to successfully calibrate the Semi-Distributed Land Use-based Runoff Process (SLURP) model in Manitoba, Canada. Solaiman & Simonovic (2010) used the NARR data in the Upper Thames River regional hydrological basin and reported satisfactory performance of such data in data-scarce regions. NARR was effectively used to calibrate the parameters of the hydrologic model in regions with limited precipitation gauges (Trubilowicz et al. 2016).

The NARR precipitation dataset, obtained from the NOAA Operational Model Archive and Distribution System (NOMADS) from 26–31 August 2011, corresponding to Hurricane Irene was used in this study as the baseline condition.

European Centre for Medium-Range Weather Forecasts (ECMWF)

The ECMWF (Molteni et al. 1996) simulates the effect of uncertainties in the initial conditions by perturbing the initial conditions of the control forecasts to obtain 50 additional perturbed ensemble members. The spatial resolution of the ECMWF ensemble members used in this study is 0.25°, while the HRES has a spatial resolution of 0.125°. ECMWF is widely used for hydrologic prediction, and many studies address the skill and advantages of ECMWF for hydrologic forecasting (Buizza 2008; Pappenberger & Buizza 2009). For example, meteorological inputs from ECMWF form the weather inputs for the European Flood Alert System (EFAS) (Thielen et al. 2009). EFAS has a post-processing scheme to correct forecast biases in the outputs and thus produce more reliable results with a reduced, but more meaningful, spread. EFAS aims at increasing preparedness for floods in trans-national European river basins by providing local water authorities with medium-range probabilistic flood forecasts.

In total, 52 precipitation fields, issued on 27 August 2011, were obtained from the Meteorological Archival and Retrieval System (MARS) for this study.

Global Ensemble Forecast System Reforecast (GEFS/R)

The GEFS is a weather forecast model comprising 21 ensemble members (Hamill et al. 2013, 2015). The GEFS accounts for the amount of uncertainty in a forecast by generating an ensemble of multiple forecasts, each minutely different, or perturbed, from the control forecast. GEFS is composed of one control forecast and 20 perturbed forecasts. The GEFS retrospective forecast dataset has a spatial resolution of 0.5° and is publicly available from 1985 to present. Reforecasts were generated using initial conditions obtained from high-quality reanalyses data using assimilation systems similar to operational systems. Hamill et al. (2015) found that reforecasts are useful for the calibration of relatively uncommon phenomena such as heavy precipitation. Saleh et al. (2016) used GEFS to forecast streamflow, visually and statistically quantifying the results of the analysis to show that GEFS forecasts can be used to produce reliable forecasts from an operational perspective. GEFS is also currently being used experimentally as a part of the MMEFS at four River Forecast Centres (RFCs) in the United States (Philpott et al. 2012). In total, 21 precipitation fields, issued on 27 August 2011, were obtained from NOMADS for the purpose of this study.

Precipitation pre-processing and resampling

The gridded precipitation data were rescaled and projected to the 2 × 2 km standard hydrologic grid (SHG) to match the physical parameter grid scale used by HEC-HMS (Maidment & Djokic 2000). The raster datasets were smoothed during the resampling process to a finer resolution from the native NARR, ECMWF, ECMWF-HRES and GEFS resolutions. Four raster resampling techniques from the QGIS ‘gdal_warp’ package (Bivand et al. 2014) were used to evaluate the importance of the resampling methods for the spatial variation of precipitation. The
resampling methods were the nearest neighbour, bilinear interpolation, cubic interpolation and cubic spline methods.

The nearest neighbour resampling method is a technique for resampling raster data in which the value of each cell in an output raster is calculated using the value of the nearest cell in an input raster. Raster datasets resampled using the nearest neighbour method showed noticeable positional errors, especially along linear features, where there was a visible realignment of pixels (Figure 2).

The bilinear resampling technique uses a weighted average of the four nearest cell centres such that the closer an
input cell centre is to the output cell centre, the higher the influence of its value on the output cell value. The raster datasets obtained using this method showed a more continuous variation in precipitation depths compared with the raster datasets obtained using the nearest neighbour resampling technique (Figure 2). Similarly, continuous variation of rainfall depths was observed in the raster datasets obtained using the cubic and the cubic spline resampling methods. As an accurate approximation of reality, it is necessary for the raster dataset to represent the continuous rainfall surface, or a smooth transition from one grid cell to another. HEC-HMS calculates the rainfall as Mean Areal Precipitation (MAP) for each time step, where MAP is the average of the rainfall in the set of cells that represent the watershed (Feldman 2000). Upon analysing the different datasets, negligible differences were found in the accumulated precipitation over the watersheds regardless of the choice of resampling techniques. The bilinear resampling method was selected as it was the default option in the framework and is widely used when interpolating precipitation forecasts from one high-resolution grid to another (Accadia et al. 2003).

**HEC-HMS model implementation**

The model used in this study is a subset of the Hudson River basin model implemented by Saleh et al. (2016). The NARR gridded precipitation data were used to force the hydrologic model, and the model was calibrated by comparing simulated flow hydrographs against hourly river flow observations to obtain representative runoff volume and peak flow (Saleh et al. 2016). The hydrologic model was initialized using archived historic data from NWIS without data assimilation schemes to update the state of the system. The Nelder-Mead optimization method was used with a root mean square objective function to produce a maximum fit between the simulated stream flows and observations (Barati 2011; Seyoum et al. 2015). Also, visual and statistical comparison was carried out with observed flows to obtain an accurate simulation of discharge at the gauging stations. For a detailed description of the implementation and validation of the ensemble-based regional hydrologic for the entire Hudson River basin, the reader may refer to Saleh et al. (2016).

The HEC-HMS model was then forced with meteorological datasets from NWP models described above under ‘Meteorological datasets and resampling methods’. We compared the precipitation and streamflow retrospective forecasts for Hurricane Irene that were issued on 27 August 2011 at 00:00, for a 96-hour forecast horizon with the precipitation and streamflow of reference obtained using NARR.

**Evaluation criteria**

The uncertainty from raw precipitation forcing and its propagation through a hydrologic streamflow prediction model is addressed in this section.

The rank histogram was used to determine the reliability of ensemble precipitation forecasts. The performance of the hydrologic model was evaluated using Nash-Sutcliffe efficiency (NSE) and bias (in %) between observed (Thielen et al. 2009) and simulated flows. Also, visual comparisons of the precipitation and flow forecasts with observations were carried out along with forecast visualization tools, namely the peak-box method (Zappa et al. 2013) and threshold exceedance diagram (Fan et al. 2014), to supplement the statistical evaluation.

**Precipitation evaluation criteria**

The precipitation forecasts were evaluated using the Talagrand rank histogram. The rank histograms are generated by repeatedly tallying the rank of the verification (usually an observation) relative to values from an ensemble sorted from lowest to highest. If an n-member ensemble and the verification are pooled into a vector at each data point and sorted from lowest to highest, then the verification is equally likely to occur in each of the n + 1 possible ranks. This process is repeated over all the sample points and the rank of the verification is tallied to obtain a histogram (Hamill 2001). Rank histograms allow us to gain a quick understanding of the ensemble spread and forecast reliability. In this context, a U-shaped histogram will mean that there is a lack of variability in the ensembles and there is a high probability that the observed value will fall outside the predicted values. On the other hand, if the observed values are always around the centre of the predicted values, the range is too broad, and an inverted U-shaped histogram would result. In contrast, a sloped histogram is indicative of consistent
biases in the forecasts (Abaza et al. 2013), while a reliable ensemble forecast is represented by a flat rank histogram (Candille & Talagrand 2005). Studies suggest that caution must be exercised when interpreting rank histograms, as credulous use can lead to errors in the interpretations of results (Hamill 2001).

Flow evaluation criteria

Two statistical criteria were selected to evaluate the streamflow forecasts, and an assessment of the forecasts’ skill was carried out using a threshold exceedance diagram. The models were statistically evaluated using NSE and bias (in %), which are used in a number of hydrologic studies. The NSE measures the fraction of the variance of the observed flows explained by the model in terms of the relative magnitude of the residual variance (‘noise’) to the variance of the flows (‘information’); the optimal value is 1.0 and values should be larger than 0.0 to indicate ‘minimally acceptable’ performance (Nash & Sutcliffe 1970; O’Connell et al. 1970).

\[
NSE = 1 - \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} (O_i - \bar{O_i})^2}
\]  

where \( N \) is the number of compared values, \( P_i \) is the simulated (forecast) value, \( O_i \) is the observed value and \( \bar{O_i} \) is the average of \( O_i \) time series.

The bias (in %) measures the average tendency of the simulated values to be larger or smaller than the observed ones. The optimal value of bias is 0.0, with low-magnitude values indicating accurate model simulation. Positive values indicate overestimation, whereas negative values indicate underestimation (Yapo et al. 1996).

\[
BIAS = \frac{100 \sum_{i=1}^{N} (P_i - O_i)}{\sum_{i=1}^{N} O_i}
\]  

Additionally, to complement the visual comparison, the peak-box method was used to obtain quantitative insights on the forecasted magnitude and time of the peaks. Such visual tools, which provide information on peak flow and time from an ensemble framework, are important where peak magnitude and times matter, for example, emergency preparedness managers to deploy flood relief measures. The ‘Peak-Box’ is a rectangle which has the lower left coordinate set to the time of the first peak flow of all members and the lowest peak magnitude of the flow members during the forecast period. The upper right coordinate is defined by the latest time of occurrence of peak and the highest discharge of all members during the forecast period. Theoretically, 25% of all the peaks of the ensemble members are included in an ‘Interquartile Range (IQR) Box’, which is a rectangle bounded by lower left coordinates from the 25% quartile of the peak timing and peak discharges and upper right coordinates from the 75% quartile of the peak timing and discharges of all members (Zappa et al. 2013).

RESULTS

Analysis of precipitation

The sub-grid scale of the sub-basins was unable to capture the spatial distribution of accumulated precipitation depths (Figure 2). For example, the highest resolution forecast (ECMWF-HRES) had only two grid cells covering the smallest sub-basin (Saddle River). Resampling to higher resolutions using the different methods did not significantly alter the spatial variation of precipitation over the three sub-basins. Similar lack of variation was observed for the coarser resolution forecasts (GEFS and ECMWF) across the intermediate and large area catchments (Figure 2).

The spread of the ensemble members in terms of the accumulated precipitation was also analysed. The analysis was based on comparing ECMWF, ECMWF-HRES and GEFS precipitation datasets with NARR. The latter was considered as the precipitation dataset of reference (Figure 3). For the Saddle River, the NARR accumulated precipitation was 140 mm (Figure 3 and Table 1). The accumulated precipitation for the GEFS ensemble members ranged from 110 to 200 mm (Figure 3 and Table 1), while ECMWF ensemble members’ accumulated precipitation ranged from 90 to 220 mm (Figure 3 and Table 1). A time lag was observed between NARR and ECMWF, with ECMWF showing a delay in the onset of precipitation by approximately 2 hours. The accumulated precipitation for ECMWF-HRES (deterministic forecast) agreed well with
NARR, with the former showing an accumulation of 150 mm compared with NARR (Figure 3 and Table 1). The control members for both ECMWF and GEFS slightly over-predicted the accumulated precipitation and showed similar precipitation accumulations (∼170 mm).

For Croton River, NARR precipitation of reference showed an accumulation of 130 mm. The accumulated precipitation ranged from 125 to 190 mm for GEFS (Figure 3 and Table 1) and 80 to 195 mm for ECMWF (Figure 3 and Table 1). However, only five out of the 21 GEFS ensemble members (corresponding to 24%) overestimated the accumulated precipitation (Figure 3 and Table 1). In this context, precipitation is seen to be well captured by the ECMWF ensembles’ uncertainty envelope. The

Table 1 | Comparison of accumulated NARR precipitation and accumulated ensemble percentiles from 21 GEFS ensemble members, 51 ECMWF ensemble members, ECMWF-HRES

<table>
<thead>
<tr>
<th>Station ID/name</th>
<th>Basin area (km²)</th>
<th>NARR</th>
<th>ECMWF-HRES</th>
<th>GEFS</th>
<th>ECMWF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5th percentile</td>
<td>Median</td>
<td>95th percentile</td>
<td>5th percentile</td>
</tr>
<tr>
<td>01391500 Saddle River at Lodi, NJ</td>
<td>140</td>
<td>120</td>
<td>169</td>
<td>197</td>
<td>101</td>
</tr>
<tr>
<td>01375000 Croton River at Croton Dam, NY</td>
<td>980</td>
<td>134</td>
<td>160</td>
<td>184</td>
<td>88</td>
</tr>
<tr>
<td>01371500 Wallkill River at Gardiner, NY</td>
<td>1,800</td>
<td>80</td>
<td>137</td>
<td>167</td>
<td>96</td>
</tr>
</tbody>
</table>
ECMWF-HRES forecast overestimated the precipitation by approximately 25 mm when compared with NARR. The control members for both ECMWF and GEFS had a tendency to overestimate the accumulated precipitation and had similar magnitudes (∼165 mm).

Finally, for Wallkill River, the accumulated precipitation from NARR was 105 mm. GEFS ensemble members showed an accumulated precipitation that ranged from 80 to 190 mm (Figure 3 and Table 1), while ECMWF ranged from 95 to 200 mm (Figure 3 and Table 1). All ECMWF ensemble members projected an accumulated precipitation that ranged from 95 to 171 mm, except for one ensemble member that was overestimating the event (Figure 3 and Table 1). The ECMWF-HRES model overestimated the accumulated precipitation by approximately 60 mm compared with the NARR reference (105 mm). The GEFS control member showed an accumulation of 135 mm, while the ECMWF control member shows an accumulation of 170 mm, corresponding to a marked overestimation when compared with NARR.

The Talagrand rank histogram showed that GEFS ensemble members tend to overestimate the precipitation, as evident from the L-shaped histogram, which indicates biased forecasts at all three sub-basins (Figure 4). Compared with the other two stations, GEFS inputs for Croton River show noticeable overestimation, since the first rank of the histogram is overloaded (Figure 4(b1)). The ECMWF inputs in both Saddle River (Figure 4(a2)) and Croton River (Figure 4(b2)) showed a high spread, corresponding to higher uncertainty in the ensemble forecasts. An ensemble that is best indicative of the verification is one where the frequency of occurrence corresponds to the equiprobable line, indicated by the dotted line in Figure 4. The general tendency, as seen from the rank histogram, was that GEFS presented biased forecasts while ECMWF ensemble members were characterized by a relatively larger spread (Figure 4).

**Analysis of discharge hydrographs**

The discharge hydrographs were analysed visually with respect to the magnitude of peak discharge, and statistically using NSE and BIAS criteria to compare the streamflow outputs. Furthermore, a threshold exceedance diagram along with the peak-box method was used to gain a better understanding of the peak discharge time of occurrence and magnitude and validate the efficiency of the operational forecast system.

For the Saddle River basin, a peak flow of 150 m$^3$/s was observed by USGS at the basin outlet (Lodi, NJ) during Hurricane Irene, which was 70 m$^3$/s above the major flood threshold of 80 m$^3$/s (Figure 5(a1) and Table 2). The peak flow ranged from 105 to 225 m$^3$/s when the model was forced with GEFS precipitation inputs (Figure 5(b1) and Table 2), while flow ranged from 70 to 250 m$^3$/s for ECMWF precipitation inputs (Figure 5(c1) and Table 2). Both ECMWF and GEFS control members showed peak flows that were overestimated by 27 and 25%, respectively, when compared with the observed flow (Figure 5(b1) and 5(c1)). ECMWF-HRES predicted the peak within 10 m$^3$/s, corresponding to 6% peak flow error (Figure 5(c1) and Table 2). The peak-box method was used as an objective quantitative criterion to quantify the uncertainty in the peak flow magnitude and timing. It showed that for the GEFS simulations, the middle 50% of the ensemble members overestimated the peaks with a slight delay in peak timing (Figure 6(a1)), while ECMWF simulations showed that the observed peaks fall within the IQR of the simulated peaks but with a marked delay in the peak timing (Figure 6(b1)). Additionally, the peak flows simulated using ECMWF meteorological inputs showed a higher range of peak timing and magnitude (Figure 6(b1)). The ECMWF-HRES simulation had an NSE of 0.72 and a bias of 4%. Forecasts across the median of all GEFS ensemble members showed a relatively higher NSE (0.67) compared with ECMWF (NSE = 0.57) ensemble members (Figure 7). The median of all GEFS members tended to be more biased (16%) compared with the median of all ECMWF ensemble members. The NSE of the median of all ensemble streamflow simulation including ECMWF-HRES was 0.72, with a bias of ∼2% (Figure 7). The use of the multiple models to simulate streamflow indicated an improvement in the NSE and an overall reduction in biases.

For Croton River on Hudson, a peak flow of 635 m$^3$/s was recorded during this event (Figure 5(a2)). Peak flows ranging from 620 to 1,220 m$^3$/s were simulated when the model was forced with GEFS inputs (Figure 5(b2)). The range of peak flows was higher for ECMWF, with the peak flows ranging from 190 to 1,355 m$^3$/s (Figure 5(c2)). Peak flows were overestimated by both GEFS and
ECMWF control members by about 60 and 55%, respectively, compared with the observed flows (Figure 5(b2) and 5(c2)). ECMWF-HRES predicted the peak more accurately compared with the control forecasts of both ensembles, although the peak error was approximately 200 m$^3$/s (Figure 5(c2)). The peak-box plots showed that the range of peak magnitude and timing from GEFS simulations was lower (Figure 6(a2)) compared with the peak magnitude and timing range from ECMWF simulations (Figure 6(b2)). The peak-box plot also depicted the
marked overestimation of the peak flow from the GEFS simulations, with the middle 50% members overestimating the peak with a slight delay in the peak timings (Figure 6(a2)). The peak-box plot from the ECMWF simulation was characterized by the observed peak falling within the IQR of the peak magnitude but outside the IQR of the
peak timings and a slight overestimation of the simulated peak median (Figure 6(b2)). Streamflow forecasts using ECMWF-HRES showed an NSE of 0.72 and a bias of 20%, while forecasts using ECMWF and GEFS showed a median NSE of 0.64 and 0.45, respectively. The median bias of GEFS ensemble members for Croton River is lower (34%) compared with the median bias of ECMWF ensemble members (−3.5%). The median of the ensembles and ECMWF-HRES showed a relatively good fit with a NSE of 0.8 and a bias of 9.6% (Figure 7).

Finally, for Wallkill River at Gardiner NY, the observed peak flow was 860 m$^3$/s (Figure 5(a3)). The peak flow ranged from 475 to 1,860 m$^3$/s when the model was forced with GEFS precipitation inputs (Figure 5(b3)) and 440 to 1,790 m$^3$/s when the model was forced with ECMWF inputs (Figure 5(c3)). Both GEFS and the ECMWF control members over-predicted the peak discharge with a peak error of 37 and 67%, respectively. The approximately similar ranges of peak magnitudes and timings are better visualized in Figure 6(a3) and 6(b3). However, consistently with the GEFS simulated peaks at the other two sub-basins, the middle 50% of the GEFS simulations overestimated the peak magnitude (Figure 6(a3)). The observed peak was within the peak magnitude defined by the IQR box in the case of ECMWF simulated peaks with a comparatively higher delay in the timings of GEFS simulated peaks (Figure 6(b3)). The ECMWF-HRES overestimated the peak flow with a percent error of 73% compared with the observed peak (Figure 5(c3)) and showed an NSE of 0.22 and a bias of 57% (Figure 7). It was noted that for this particular event, the ECMWF-HRES data had a general tendency to overestimate the rainfall at the three sub-basins, and this was particularly significant for the larger sub-basin. In the case of the Wallkill River, the accumulated rainfall from ECMWF-HRES was overestimated by approximately 60 mm compared with the NARR rainfall (105 mm). This over-prediction in rainfall translated to positive bias in the streamflow simulated by the hydrologic model. However, it cannot be concluded from this event-based case alone whether it is a general trait of ECMWF-HRES to over-predict the rainfall at larger sub-basins.

The median of all ensemble members from the two models was representative of the observed flow and showed an NSE of 0.86 and bias of 6.2% (Figures 6 and 7).

The threshold exceedance persistence diagram was used to better understand the forecast persistence from an...
operational management perspective (Thielen et al. 2009; Fan et al. 2014; Thiemig et al. 2015; Saleh et al. 2016). We compared a single forecast issued on August 27 2011 using two different NWP ensemble models applied to three catchments. A higher probability of exceedance of the threshold, which compares correctly with the observations, indicates a more skilful forecast. A probability of 100% indicates that all members are projected to exceed a defined threshold (Figure 8). The major flood threshold was used at Saddle River at Lodi, while the observed peak flow during Hurricane Irene was used as a threshold for the other two sub-basins (Figure 8). A higher percentage of GEFS ensemble members projected an exceedance of the threshold compared with the ECMWF ensemble members at all three stations for the time periods of interest (when observation exceeds the threshold). This corresponds to more GEFS ensemble members exceeding the threshold compared with the ECMWF ensemble members. Approximately 50% of both GEFS and ECMWF ensemble members predicted a major event after the flood had receded (Figure 8). For this event, ECMWF-HRES showed a slight lag in predicting the event for all three stations, which further reiterates the benefit of using probabilistic forecasts for decision-making.

**SUMMARY AND DISCUSSION**

The objective of this study was to evaluate the uncertainty in streamflow forecasts inherited from ensemble-based NWP meteorological inputs. This research contributes to the
growing literature that shows the promising capabilities of multi-model systems to better describe the uncertainty in streamflow predictions.

Streamflow in selected sub-basins of the Hudson River basin was simulated using a hydrologic model forced with 73 ensemble precipitation members from NOAA/ESRL GEFS, ECMWF and ECMWF-HRES. A retrospective forecast of an extreme flood event, Hurricane Irene, was used to evaluate the uncertainty in three sub-basins. The meteorological datasets from the NWP models were resampled from the native grid resolution to the SHG using four different resampling methods, namely the nearest neighbour, bilinear, cubic and cubic spline methods. The datasets obtained from different resampling methods were analysed and showed negligible differences in terms of the accumulated precipitation depths over the three sub-basins (Figure 2). The negligible differences are attributed to the sub-grid scale of the watersheds, which was unable to capture the spatial variation of precipitation, and the semi-distributed hydrologic model, which averaged the precipitation over the sub-basins (Figure 2). The bilinear resampling method was selected after analysing the four different methods.

The uncertainties in retrospective reforecasts were evaluated using statistical metrics and visual comparisons. The approach was based on comparing the time series of accumulated precipitation between the reanalysis dataset (NARR) and inputs obtained from ensemble-based NWP models. An overall assessment of the reliability of the accumulated precipitation, using the Talagrand rank histogram, showed that GEFS members were inclined to be biased, while ECMWF members were characterized by a higher spread (Figure 4). Even though a highly spread uncertainty envelope better captures a given flood event, it is important to establish a confidence limit (reliability) that can be conclusively used to inform decision-making. In particular, higher spread in forecasts hinders efficient and effective operational decision-making compared with forecasts that are sharp and reliable. This bolsters the need for pre-processing the precipitation inputs to reduce the systematic biases and reliably quantify uncertainty using different probabilistic methods (Robertson et al. 2015; Verkade et al. 2015; Roulin & Vannitsem 2015) before forcing it in the hydrologic model.

The work demonstrated the significant dependency of streamflow forecasts on meteorological inputs in the case of this extreme rainfall event. It showed how uncertainties from such inputs propagate in the hydrologic model and reflect on the simulated streamflow. For instance, a temporal

---

**Table 1**

<table>
<thead>
<tr>
<th>Station &amp; flood threshold</th>
<th>Time</th>
<th>Forecast</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saddle River at Lodi, NJ (01391500) 80 m3/s</td>
<td>8/27/2011</td>
<td>HRES: 0 0 0 0 0 0 0 100 100 100 100 0 0 0 0</td>
<td>ECMWF: 0 3 53 97 97 95 77 36 3 0 0 0</td>
</tr>
<tr>
<td>Croton Dam, NY (01375000) 567 m3/s</td>
<td>8/28/2011</td>
<td>HRES: 0 0 0 0 0 0 0 10 90 100 100 100 90 48 0 0 0 0</td>
<td>ECMWF: 0 10 86 100 100 76 29 0 0 0 0</td>
</tr>
<tr>
<td>Wallkill River at Gardiner, NY (01371500) 850 m3/s</td>
<td>8/29/2011</td>
<td>HRES: 0 0 0 0 0 0 0 0 0 0 100 100 100 100 0 0 0 0</td>
<td>ECMWF: 0 0 0 0 0 0 0 0 0 0 43 63 59 35 14 2 2 0 0</td>
</tr>
<tr>
<td></td>
<td>8/30/2011</td>
<td>HRES: 0 0 0 0 0 0 0 0 0 0 14 81 81 71 52 24 10 5 5 0</td>
<td>ECMWF: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
</tbody>
</table>

**Figure 1** Colour-coded threshold exceedance diagram for Hurricane Irene forecasts at 6-hour intervals using the flood threshold for each USGS station. The flow values below the station ID are the flood threshold for the stations. The major flood threshold used for Saddle River at Lodi was obtained from the NOAA National Weather Service major flood thresholds. Peak discharge values from Hurricane Irene were used as thresholds for both Croton River and Wallkill River. The 96-h forecast horizon for each model from the simulation date shown in the left column is represented by the x-axis. The time period in which the major event hydrological record exceeds the equivalent alert threshold is indicated using a dark red cell, while the cell values refer to the percentage of ensemble members that were projecting a major flood. For instance, if the value is 100, then all flow ensemble members of the particular ensemble are projecting the event within a given time interval, while 0 means none of the ensemble members are exceeding the defined event threshold. The observed occurrence of the threshold is exhibited at the last row for each station, the red colour-code indicates an observed flow higher than the defined flood threshold, while the green cell is flow below the major threshold. Reported time is in UTC. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/nh.2018.182.
delay in the onset of precipitation translated to a lag in the prediction of peak discharge. This may have consequences for issuing timely flood alerts, which was also exhibited in the threshold exceedance diagram (Figure 8). The latter showed the temporal evolution of flow forecasts exceeding the defined flood thresholds with a number of ensemble members projecting flooding even after the event. Consequently, caution needs to be exercised when interpreting the threshold exceedance diagram (Sankarasubramanian et al. 2009; Anghileri et al. 2016). The median of all 73 ensemble streamflow forecasts improved the NSE and reduced biases in streamflow across the different sub-basins compared with only using the median from one NWP model. However, care should be exercised when only using the median for operational decision-making without an associated uncertainty envelope.

While this study only accounts for the total uncertainty in the ensemble streamflow forecasts, it offers perspectives for future efforts and enhancements to account for both meteorological and hydrologic uncertainties. For example, separate approaches could be used to first model the forcing input uncertainty (using pre-processing techniques) and then the hydrologic uncertainty (using post-processing techniques). This necessitates methodologies to incorporate and evaluate future enhancements such as pre-processing and post-processing techniques for specific applications and study areas (Kang et al. 2010; Zalachori et al. 2012). There is a need for efforts focused on additional case studies directed toward extreme events using available meteorological datasets such as TIGGE (Bao et al. 2011). Outcomes from such analogous efforts along with comprehensive forecast verification strategies are potentially useful for operational forecasters and support more informed decisions during extreme events. In addition to using an ensemble of meteorological inputs to account for atmospheric uncertainty, modelling approaches could also be used to quantify the hydrologic uncertainty by executing a given hydrologic model with a different set of parameters to account for the errors arising from initial conditions (e.g. soil moisture and snow conditions) (Zappa et al. 2008; Dong et al. 2015; Kumar et al. 2015). A multi-hydrologic modelling approach could also be used by testing various hydrologic models to account for the model structure uncertainty (e.g. lumped, semi-distributed, distributed) (Liang et al. 1994; Gochis et al. 2013). Along with the scientific challenges, there are several institutional and communicative challenges in applying EPS to operational flood forecasting, in particular the communication of probabilistic forecasts to the general public (Morss et al. 2005; Faulkner et al. 2007). The outcome of this work offers important perspectives for operational frameworks running multi-model ensembles and shows that multi-model ensembles are useful to better describe the uncertainty in hydrological forecasts.

ACKNOWLEDGEMENTS

This work was funded by a research task agreement entered between the Trustees of the Stevens Institute of Technology and the Port Authority of New York and New Jersey, effective 19 August 2014. We would like to thank the two anonymous reviewers who reviewed this manuscript. Their comments, suggestions and valuable insights led to significant improvement in the quality of our manuscript.

REFERENCES


Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A. & Speranza, A. 2005 Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. Weather Forecast. 18, 918–932.


Amengual, A., Romero, R., Vich, M. & Alonso, S. 2009 Inclusion of potential vorticity uncertainties into a...


Bivand, R., Keitt, T. & Rowlingson, B. 2014 *rgdal: Bindings for the Geospatial Data Abstraction Library.* R package version 0.8-16.


First received 17 November 2017; accepted in revised form 16 July 2018. Available online 20 August 2018