

The Effect of Large Scale Atmospheric Uncertainty on Streamflow Predictability

Dingchen Hou², Kenneth Mitchell¹, Zoltan Toth¹,

Dag Lohmann³ and Helin Wei²

¹ Environmental Modeling Center, NCEP/NWS/NOAA, Camp Springs, MD, USA

² SAIC at EMC/NCEP/NWS/NOAA, Camp Springs, MD, USA

³ Risk Management Solution, Ltd., London, UK

September 30, 2008

Corresponding author:

Dingchen Hou

EMC/NCEP/NOAA

W/NP2 NOAA WWB #207

5200 Auth Road

Camp Springs, MD 20746

(301)7638000 ext7015

Dingchen.Hou@noaa.gov

Abstract

Hydrological processes are strongly coupled with atmospheric processes related for example to precipitation and temperature and a coupled atmosphere-land surface system is required for a meaningful hydrological forecast. Since the atmosphere is a chaotic system with limited predictability, ensemble forecasts offer a practical tool to predict the future state of the coupled system in a probabilistic fashion, potentially leading to more complete and informative hydrologic prediction. As ensemble forecast with coupled meteorological-hydrological models are operationally running at major numerical weather prediction centers, it is currently possible to produce gridded streamflow prognosis in the form of probabilistic forecast based on ensembles. Evaluation and improvement of such products require comprehensive assessment of both components of the coupled system.

In this article, the atmospheric component of a coupled ensemble forecasting system is evaluated in terms of its ability to provide reasonable forcing to the hydrological component and the effect of the uncertainty represented in the atmospheric ensemble system on the predictability of streamflow as a hydrological variable. The Global Ensemble Forecast System (GEFS) of NCEP is evaluated following a “perfect hydrology” approach, in which its hydrological component, including the Noah Land Surface Model and attached river routing model, is considered free of errors and the initial conditions in the hydrological variables are assumed accurate. The evaluation is performed over the Continental United States (CONUS) domain for various sizes of river basins. The results from the experiment suggest that the coupled system is capable of generating useful gridded streamflow forecast when the land surface model and the river routing model can successfully simulate the hydrological processes, and the ensemble strategy significantly improve the forecast. The expected forecast skill increases with increasing size of the river basin. With the current GEFS system, positive skill in short range (1 to 3 days) predictions can be expected for all significant river basins, and for the major rivers

with mean streamflow over $500 \text{ m}^3 \text{ s}^{-1}$, significant skill can be expected from extended range (the second week) predictions. Possible causes for the loss of skills, including the existence of systematic error and insufficient ensemble spread, are discussed and possible approaches for the improvement of the atmospheric ensemble forecast system are also proposed.

1. Introduction

Flooding and drought are the most frequent natural hazards and water resource management is one of the most challenging problems the world is facing. Therefore, hydrological forecast, especially streamflow forecast, is of great interest and it is a major application of numerical weather prediction (NWP) output. NWP forecasts of precipitation and temperature can be incorporated into a flood warning system and the forecast lead time can be significantly increased (e.g. Gourley and Vieux, 2005 and Krzysztofowicz, 2002). While calibrated NWP products are used to force hydrological models in traditional methods (e.g. Werner et al., 2005), one emerging approach is to use a coupled meteorological-hydrological modeling system (e.g. Verbunt et al. 2006).

Since precipitation forecasts exhibit large uncertainties and many hydrological services consider the use of forecast precipitation to introduce an unacceptable degree of uncertainty into their forecasts and make the decision making system problematic, hydrologic forecasts must be framed in a probabilistic form. The limitation of deterministic NWP products can be addressed by ensemble prediction systems (EPS) which incorporate uncertainties in the initial conditions and other factors of the modeling process and provide multiple weather forecast (e.g., Houtekamer et al., 1996; Buizza, 1997; Houtekamer and Mitchell, 1997; Toth and Kalnay, 1997). It is generally accepted that to capture case dependent variations in forecast uncertainty, one must follow an ensemble approach. While various techniques, such as Ensemble Pre-Processor, are used to re-generate ensemble members, direct use of the NWP ensemble members (Pappenberger et al., 2005 and Pappenberger et al., 2008) provides another alternative approach. It has the advantage of facilitating the coupled meteorological-hydrological modeling.

Traditionally, hydrological forecast is made for individual river basins and the predicted streamflow is valid at the outlet of the river basin. However, a seamless environmental forecast suite requires streamflow forecast for all river basins and at any point along each river. Therefore, the most convenient approach is to provide streamflow prognosis for a mesh of grid points, just like what is done for NWP products such as precipitation and temperature. As the land surface models are greatly improved during the last decade and they are coupled with the atmospheric models in operational NWP systems (Mitchell et al., 2005) and various river routing models are developed, it is possible now to generate gridded streamflow prognosis as an NWP product to provide guidance to river forecast centers. Nevertheless, such product is not operationally available and its quality remains to be evaluated.

There are fundamental differences between the two components of the coupled modeling system in the nature of the model, the ensemble technology, and the quality of the input data. The meteorological component of the coupled system, i.e. an NWP model, is based on dynamic and physical principles. The initial conditions used to start the integration are provided by a well defined global observational network and a complicated and comprehensive objective analysis scheme. Therefore, these models are relatively mature and the accuracy of forecast, especially at short range, is high. At longer lead time, the impact of uncertainties such as in the initial conditions will increase, but the EPS approach helps to mitigate the problem. Incorporation of model related uncertainty with stochastic parameterization, stochastic perturbations and multi-model ensembles can further improve the forecast and increase the lead time (Buizza et al., 1999; Hou et al., 2008). On the other hand, the hydrological component of a coupled system, i.e. the rainfall-runoff model and the attached river routing model, despite the rapid progress in recent years, are still subject to extreme complexity of the physical process, insufficient resolution in the representation of the land surfaces characteristics, and lack of a continuous and reliable observation network. Mature techniques to generate ensemble members or quantify the uncertainties associated with the initial conditions and the model, are also

non-existing or still in its infant stage. In many cases, an ensemble of meteorological forecasts is used to drive the same hydrological integration, with the same model and from the same initial conditions. For example, Pappenberger et al. (2008) propagated each weather forecast ensemble member through the LISFLOOD model. Therefore, the development and improvement of a coupled meteorological-hydrological system is still a two-tiered task. While the hydrological community is striving to improve their models, expand the observation networks, explore various analysis methodologies, and develop ensemble strategies, the NWP community needs to evaluate and improve the ensemble forecasting system (including model, analysis and ensemble strategy) from the hydrological perspective.

The operational Global Ensemble Forecast System (GEFS) of National Centers for Environmental Prediction (NCEP), in which the atmospheric model, Global Forecasting System (GFS), is coupled with Noah land surface model (Mitchell et al., 2005), provides a good opportunity to study the feasibility of gridded ensemble streamflow forecast and evaluate the quality of the meteorological output from the perspective of hydrological forecast. This article is resulted from the first step of a study in this direction and the primary attention is focused on streamflow as the model output. In specific, the authors try to answer the following questions: (1) is the quality of the coupled system sufficient to generate any useful streamflow forecast at various lead times? (2) How is the skill of streamflow forecast dependent on the river basin size and lead time? (3) Is the uncertainty represented in the ensemble generation strategy helpful to improve the streamflow forecast, and is it sufficient to account for the uncertainties in the hydrological output? (4) What measures should be taken in modifying the atmospheric modeling component to improve the ensemble streamflow forecasting? In section 2, the coupled atmosphere-land forecast system and the ensemble prediction system used in GEFS are briefly described and the design of experiment is presented. The methodology of evaluation is discussed and the forecast verification scores used for the evaluation are reviewed in section 3, and the results are presented in section 4. Finally, a summary is provided and the implication of the results in operational forecast is discussed in section 5.

2. Model configuration and experimental design

During the last decade, the understanding of land surface processes experienced rapid progress and various land surface models (LSM) were developed (Mitchell et al., 2004) and coupled with atmospheric models in operational Numerical Weather Prediction (NWP) systems. At NCEP, the Noah LSM was implemented into the Global Forecast System (GFS) in 2004 (Mitchell et al., 2005). Both the NCEP operational global deterministic forecast (GFS) and the Global Ensemble Forecast System (GEFS) are based on this coupled GFS-Noah modeling system. These forecasts provide in their output a number of hydrological variables, including precipitation, surface temperature and runoff in each grid cell (note that the word “runoff” in this paper refers to the surface and sub-surfaces runoffs before it reaches a river channel). However, streamflow, a more useful hydrological variable which can be generated by river routing models from runoff, is not available.

River routing experiments are carried out at NCEP within the North America Land Data Assimilation System (NLDAS) project (Lohmann et al., 2004). NLDAS project (Mitchell et al., 2004) runs four land surface models at analysis mode over the Continental United States (CONUS) domain with 1/8 degree grid separation, by taking meteorological input from regional re-analysis, which includes estimated real hourly precipitation based on observations. The backbone of the NLDAS precipitation is the 1/8 degree gauge-only daily data prepared by Climate Prediction Center (CPC) of the National Oceanic and Atmospheric Administration (NOAA). This daily analysis is temporally disaggregated to hourly by applying hourly weights derived from hourly 4-km radar-based precipitation field. The project has generated retrospective analysis for the past 30 years and running daily at quasi-operational mode. Its products provide hourly value of land surface variables, including soil moisture, temperature, as well as surface and sub-surface runoffs. As observed precipitation is used in the NLDAS simulations, these products are actually in analysis mode.

The river routing model used by Lohmann et al. (2004) calculates the timing of the runoff reaching the outlet of a grid box, as well as the transport of the water through the river network. Both within grid cell and river routing time delays are represented using linear, time-invariant and causal models, which are represented by nonnegative impulse-response functions. The river flow direction mask used is a D8 model which assumes that water can leave a grid cell only in one of the eight neighboring grid cells. In terms of model structure, the river routing model has “surface water” as its state variable, which is a reservoir filled up by within grid runoff and transport from upstream grid cells, and drained off by the transport to its downstream cell. Runoff is an input or external forcing and streamflow is calculated as a diagnostic variable. Again, this NLDAS product is referred as streamflow analysis in this article. Lohmann et al. (2004) showed that this type of streamflow analysis is in general comparable to river flow observations, although significant intermodal difference exists in its daily variations. The difference is attributed to the large disparity among the LSMs, or model uncertainty, in various physical processes such as canopy conductance, aerodynamic conductance, soil moisture storage and snowmelt (Mitchell et al., 2004 and Lohmann et al., 2004).

As the river routing model provides feedback to neither the land surface model nor the atmospheric model, it can be simply attached to any land surface model to form a complete hydrological model. When it is attached to a coupled land-atmosphere forecast system or driven by forecast runoff, it will generate streamflow forecast. In this study, the river routing model is attached to the coupled GFS-Noah forecasting system and streamflow forecast of up to 16 days is generated. However, due to the retrospective nature of the experiment, it is referred as streamflow simulation. The “surface water” analysis, generated by the river routing in the NLDAS streamflow analysis as just described, is used to initialize these streamflow simulations. Therefore, the simulated streamflow at the initial time is the same as the NLDAS analysis.

The period selected for the experiment is 1 April to 30 May 2006 and the simulation is initialized at 0000 UTC each day. During this two month period, the Global Forecast System (GFS)

deterministic (single) operational forecast runs at T382 (truncated to T190 at 180h) horizontal resolution with 64 vertical levels (L64) and is referred as GFS or GFS high resolution simulation. The Global Ensemble Forecast System (GEFS) runs at T126 and L28 resolution and has a control (initial conditions are the same as GFS except for lower resolution) and 10 perturbed members (the initial conditions are slightly different from the control, generated by Breeding). The 10 perturbed members are collectively referred as GEFS and the control as CONTROL. Runoff forecasts from GFS, CONTROL and from each of the 10 GEFS members are used to force the river routing model. Although the focus is the ensemble simulations, the two single runs are included for the purpose of comparison. Therefore, there are 12 streamflow simulations up to 16 days. The operational GFS/GEFS forecast provided runoff output at 1 degree by 1 degree global grid every 6 hours from 0 to 180 hours of lead time, and 2.5 by 2.5 degrees every 12 hours from day 8 to day 16. To match the coarse resolutions of the runoff to the fine grid mesh of the river routing model and the river flow direction network, the forecast runoff is interpolated to the fine grid mesh of 1/8 times 1/8 degree over the CONUS domain. Effort is made to conserve the water volume in each coarse grid cell when the interpolation is made from the coarse grid to the fine one, but downscaling is not considered. Given the horizontal resolution of the forecast model, only the large scale atmospheric circulation patterns and the associated uncertainty are considered in this study.

3. Methodology of Analysis

To verify and calibrate a streamflow forecast, one needs to compare it with river stage observations. This comparison is often complicated by the water management at reservoirs and dams, which is not included in the modeling system. Some authors made the comparison for selected river basins where human management can be neglected (e.g. Lohmann et al., 2004), or with so-called natural flow, which is the observation modified by taking the human management into account. For a

gridded streamflow forecast, either approach involves tremendous effort to prepare the data base which is currently unavailable. As the current study is only aimed at evaluating the performance of the GEFS ensemble forecast system in term of its hydrological output, or its ability to provide a reasonable meteorological forcing to the hydrological model, this type of forecast verification is not performed.

Because the performance of a streamflow forecast is affected by both components of the coupled meteorological-hydrological system and a major part of the forecast uncertainty is from the land surface model and the river routing model, it is necessary to isolate the effects of the atmospheric model, as well as the initial conditions and ensemble strategy associated with the atmospheric variables. One way to achieve this is to take an approach similar to the “perfect model” approach in predictability studies. In this context, the hydrological model (land surface and river routing models) is assumed to be perfect and the initial conditions of the hydrological variables (soil variables and surface water) are assumed accurate. Therefore, the simulated streamflow can be evaluated by comparing it with the streamflow analysis. This approach will be referred as “perfect hydrology”. Intuitively, as both the simulation and analysis of streamflow are generated by the same modeling system (land surface and river routing models) from the same initial conditions of soil model variables and the surface water, their difference is mainly the reflection of the differences in the meteorological forcing to the hydrological component of the coupled system. As the runoff and streamflow mainly reflect the hydrological model response to the precipitation and surface temperature, the comparison is, to some extent, between forecast and observations of these two variables. In other words, by assuming that the streamflow analysis represents the truth, the comparison between analyzed and simulated streamflow will provide hints on how to improve the simulation by improving the meteorological component of the coupled modeling system. Improvement in the atmospheric ensemble system following this approach will work with any reasonable land surface and river routing model that will be available in the future. It has been noted that the Noah model used in the GFS/GEFS is slightly different from that used in NLDAS, mainly in the horizontal resolution. As the uncertainty in steamflow is dominated by

uncertainty in precipitation (Milly and Dunne 2002a, 2002b), this difference is considered to be negligible compared with that in the spatial variation of precipitation between the observation and the forecasts.

For the coupled GEFS-Noah-River system, the uncertainty considered in the simulation is only that associated with the initial conditions of the atmospheric variables. The uncertainty associated with the atmospheric model was not presented in the ensemble system although effort is in progress to implement a stochastic perturbation scheme (Hou et al., 2008). The comparison of the GEFS based ensemble streamflow simulations with that forced by a single forecast, GFS or CONTROL, with the analysis as the truth, can help to evaluate the effect of the uncertainty in the initial conditions on the quality of the simulation and necessity to include other sources of uncertainties. As far as the hydrological component is concerned, the only uncertainties considered are those associated with the meteorological forcing, i.e. precipitation, near-surface temperature and other related variables. In other words, perfect initial conditions and perfect model are assumed in the Noah-River sub-system.

To compare gridded streamflow simulations against analysis, it is important to note the special characteristics of the variable in contrast to the meteorological quantities, such as precipitation and near-surface temperature. Each grid point in the grid mesh may contain a major river, a secondary river, a minor stream or no river channel at all, depending on the area of its upstream catchment. The runoff generated in different parts of the catchment has different travel time to reach the grid cell in consideration. For any grid cell in the lower reach of a major river, the travel time from any upstream grid cell can be inferred from Fig. 1, the map of the travel time to the outlet (to Oceans or the Great Lakes) with the help of river network information. In Fig. 1, the catchment and travel time are also schematically illustrated for the grid cell of Vicksburg, MS, at the lower reach of the Mississippi River. Unlike small river basins where predictability of streamflow mainly reflects the local predictability of precipitation with similar lead time, the forecast skill of streamflow for this large river station is the combination of meteorological predictability in different parts of its catchment and at a range of lead

times. Consequently, it is necessary to test if and how the forecast skill is dependent on the river basin size.

The temporal average of streamflow at a grid point is used as a measure for the size of the river basin, or the area of the catchment for the channel outlet of the grid point. The grid points are grouped in to 20 categories based on the streamflow averaged over the two month experiment period, as shown in Table 1. This grouping is more or less arbitrary but effort is made to make each category to contain roughly equal number of grid cells, except for the lowest categories, which contain much more grid points than the higher categories. Note that the maximum of mean streamflow is over $15000 \text{ m}^3 \text{ s}^{-1}$ in the lower reach of the Mississippi and the grid points in the two lowest categories have their catchments too small compared with the resolution of the atmospheric model. In Table 2 is listed related information of a number of grid points, which will be used as examples of river basins in the analysis. In contrast to the huge river basin of Mississippi River, Vicksburg MS, the catchment areas of the Potomac River, Washington DC and the Merrimack-Concord River, Lowell MA are about the size of the grid cell of the atmospheric model and they are identified as medium sized river basins. It should be point out that the mean streamflow ($499 \text{ m}^3 \text{ s}^{-1}$) for the Merrimack-Concord River is much higher than its multi-year average (less than $200 \text{ m}^3 \text{ s}^{-1}$) due to the historic flood in mid-May 2006. The Nehalem River, FOSS OR, with its catchment area about 2000 km^2 , at least one order of magnitude smaller than the atmospheric model's grid cell, is selected as an example of small river basins. The locations of the four grid cells are marked in Fig. 2, with the catchments for the Mississippi and Potomac rivers schematically outlined.

Simulated streamflow is evaluated in terms of both deterministic (single value) and probabilistic forecast based on ensembles. For a deterministic forecast, the comparison between the analysis and a simulation presented in this study is concentrated on the time series for each grid point, and the verification scores are averaged over each category of grid points. For the time series of a

simulation s and the analysis a , spanning the same period with a length n , a commonly used verification score is the correlation coefficient,

$$C = \frac{\sum_{i=1}^n (s_i - \bar{s})(a_i - \bar{a})}{\left[\sum_{i=1}^n (s_i - \bar{s})^2 \sum_{i=1}^n (a_i - \bar{a})^2 \right]^{1/2}} .$$

C is a relative fitting measure between the two time series postulating a linear regression between them. It is also referred as coefficient of determination by some authors (e.g., Koutsoyiannis et al., 2007). Although a higher correlation does not mean better match between the two time series in terms of minimum difference, it does indicate a good match after simple bias correction or linear regression. Therefore a higher value in C indicates a potential forecast skill.

Nash-Sutcliffe Efficiency Coefficient, or NSEC (Nash and Sutcliffe 1970; Moriasi et al., 2007) is widely used to assess the predictive power of hydrological models. Defined as

$$E = 1.0 - \frac{\sum_{i=1}^n (s_i - a_i)^2}{\sum_{i=1}^n (a_i - \bar{a})^2}$$

NSEC ranges from $-\infty$ to 1 and can be interpreted as the skill score for Mean Square Error (MSE) with the analyzed (observed) mean as the reference forecast. A positive value of E indicates a better match of the two time series and represents a real forecast skill. $E=1$ corresponds to a perfect match between simulated (forecasted) and analyzed (observed) streamflow while $E=0$ indicates that the simulation is just as comparable to a “climate” forecast using the temporal mean of the analysis.

For an ensemble of simulations of continuous variable, such as streamflow, the cumulative distribution function of the predicted quantity S , can be estimated and denoted by the probability for $F(s)=p(S<s)$. Following Toth et al. (2003), the Continuous Ranked Probability Score (CRPS) can be calculated as

$$CRPS = E \left(\int_{-\infty}^{\infty} [F(s) - H(s-a)]^2 ds \right)$$

where $H(s-a)$ is the Heaviside function that takes the value 0 when $s-a < 0$, and 1 otherwise. It also can be expressed as a skill score, Continuous Ranked Probability Skill Score (CRPSS):

$$CRPSS = 1 - \frac{CRPS}{CRPS_{ref}}$$

In this formula $CRPS_{ref}$ is the CRPS of a reference forecast, which is used as a benchmark for a comparison. The reference forecast used in this paper is persistent forecast, in which the initial streamflow is taken as the forecast for all lead times. One advantage of CRPS is that it can be applied to a deterministic (single valued) simulation of s . At this case, it is reduced to the Mean Absolute Error $AE=E(|s-a|)$ when averaged over a number of cases and/or grid points.

4. Evaluation of the simulations

This study is focused on the overall performance of the coupled system in simulating the streamflow, including its magnitude and temporal evolution. Specific forecast, e.g. the time of peak flow and total discharge is possible only when the overall performance is of practical value. Therefore, primary effort is devoted to the statistics, or the temporal average, of the verification scores of the streamflow simulations. Nevertheless, an inspection of individual cases (i.e., simulations of a particular lead time and/or initialized at a particular day) is a necessary step of the study and it is helpful to the quantitative evaluation of the coupled modeling system.

a. Case studies

We start with an overview of the simulations by inspecting some randomly selected cases for specified lead times. Fig.2 shows the ensemble mean simulated streamflow with 12 day lead time, initialized at 0000 UTC 1 April 2006, the corresponding analysis, their difference (or error) and the

ensemble spread. Both the analysis and the simulation correctly depicted the major rivers and no major difference in the river network is found without looking at the error map. Both overestimation and underestimation exist and they tend to show geographic patterns, suggesting that they may be related to weather and climate patterns. However, this is not the focus of this study and should be left for further investigations. In general, larger errors are associated with larger rivers. Noting that the shading scale in the error and spread maps is one order of magnitude smaller than that in the streamflow maps, one can find that the relative error is acceptable. On the other hand, the ensemble spread also has similar pattern as the streamflow itself, and it has the same order of magnitude as the errors. An inspection for other lead times and (not shown) revealed the same characteristics, except that the magnitude of error and spread increases with lead time. These characteristics are also seen from other forecast cases (initialized from different days) and suggest that the coupled Atmosphere-Land-River system is able to generate streamflow simulations comparable to the analysis and the uncertainty in the atmospheric forcing from GEFS is also reasonable.

Fig. 3 shows two examples of the simulation trajectories, or simulated streamflow as a function of lead time, for the grid cell in which the Potomac River at Washington DC is located. The top panel is the simulation initialized at 0000 UTC 1 April. The analysis (red curve) suggests that the river is relatively calm during the first half of April, with only two noticeable events of river stage increases, on 7 and 12 April. While the CONTROL and most of the GEFS members failed to catch either event, four ensemble members do hit one of them with various amplitude and timing. Another example shows the simulations started at 0000 UTC 4 May 2006 (lower panel). Half of the members of GEFS suggested a significant flood event during the second week of the 16 day period of the integration, leading to a moderate peak value of $150 \text{ m}^3 \text{ s}^{-1}$ in the ensemble mean streamflow on 15 May. The analysis recorded a peak flow on the same day, with its magnitude ($160 \text{ m}^3 \text{ s}^{-1}$) only slightly higher. On the other hand, the GFS and CONTROL simulations delayed the peak by two days and GFS tripled the peak value. These examples suggest that the ensemble strategy applied to the atmospheric component

of the coupled system is effective in representing the atmospheric uncertainty and helps to improve streamflow simulations.

Two time series of simulations for a selected grid cell and with a fixed lead time, taken from the daily simulations during the two month period, are shown in Figs. 4 and 5. For the lower Mississippi River at Vicksburg, MS, which represents the largest basin over CONUS with mean streamflow of $14788 \text{ m}^3 \text{ s}^{-1}$, the major event of increased streamflow at the end of April and a number of minor events are well predicted 15 days in advance by all single simulations (Fig. 4). Also, for most days, the analysis is embraced by the ensemble numbers. For the Merrimack-Concord River at Lowell, MA (mean streamflow $499 \text{ m}^3 \text{ s}^{-1}$), the flood in mid-May, one of the most significant flood in recorded history of the river, is predicted in 5-day forecasts (Fig. 5), suggesting that the system has skills in forecasting extreme events. During the low flow period, the ensemble spread tends to be insufficient and all simulations are close to each other but significantly different from the analysis, indicating significant systematic error, that is probably due to over-prediction of light rain event, a common feature of NWP models.

These cases show that the coupled streamflow simulations, especially those with the ensemble approach, do have forecast skills and the skills are strongly dependent on the river basin size and lead time, and the findings will be confirmed by quantitative evaluation in the following subsections.

b. Deterministic Forecast skills

From the time series in Figs. 4 and 5 it can be seen that the temporal correlation between the simulations and analysis is high. To show this in a quantitative measure for all lead times and all ranges of river sizes, the correlation coefficient is calculated for each grid point and each lead time, and the results are displayed for two grid cells in Fig. 6.

The Potomac River at Washington DC is a typical example of medium sized rivers. The correlation coefficients for all of the simulations are very close to unit during the first two days, a

reflection of the fact that all members are initialized with the same initial fields of the surface water, which is the same as the analysis, and the high accuracy of short range precipitation prognosis. Starting from day 3, the correlation gradually decreases as lead time increases, to about 0.5 at day 5-7, and 0 at about day 10. This is not a surprise if one notes that the travel time (Fig. 1) even at its headwater is less than 10 days. These results suggest that the simulations have significant potential forecast skill at least for the first week. It is also noted that the GEFS ensemble mean outperforms most of the GEFS members at most lead times, as expected from the properties of EPS strategies.

For small river basins, faster decrease of correlation is expected as the catchment is small and the travel time is shorter. However, many grid cells representing small river basins in the Pacific West show opposite results. Depicted in the lower panel of Fig. 6 is for the Nehalem River, at Foss OR. This river station, with its mean streamflow of about $40 \text{ m}^3 \text{ s}^{-1}$, had excellent agreement between the analysis and the USGS river gauge observations (see Lohmann et al., 2004). The simulation-analysis correlation is higher than 0.8 up to day 12. Higher predictability has been noticed in precipitation verification over the West region and may be related to the characteristic of weather patterns and local effects. As the focus of this article is the general assessment of the coupled meteorological-hydrological modeling system with large scale atmospheric forcing, this phenomenon is not studied in detail.

To reveal the general trend in the variation of correlation coefficient with lead time and river basin size, the score is averaged for each of the 20 categories of grid points over the two month period (see Fig.7). The CONTROL simulation (the upper right panel) has positive correlation with analysis, for all ranges of river size and all lead times. For the large rivers (categories 18, 19 and 20, mean stream flow over $500 \text{ m}^3 \text{ s}^{-1}$), the correlation is higher than 0.5 up to two weeks of lead time while it reduces to 0.4 at day two for the smallest streams. For the medium sized rivers (categories 11-17, mean streamflow $55\text{-}500 \text{ m}^3 \text{ s}^{-1}$), This useful lead time decreases from 10 day for category 17 to seven day for category 11. Significant correlation (higher than 0.5) only last four days for small rivers of category

3 to 10 (mean streamflow of 10-55 m³ s⁻¹). The smooth variation of correlation with river categories suggests that the river size grouping described in section 3 is reasonable. It is interesting to note that the GFS simulation, despite its higher resolution, is not as good as CONTROL in terms of correlation coefficient (upper left). Neither the mean score of the 10 GEFS members can match the CONTROL simulation (lower left). However, a major improvement is achieved by taking the GEFS ensemble mean simulation (lower right), especially for the second week forecasts over small and medium sized basins. This indicates that the EPS strategy is effective in improving streamflow simulations.

Fig. 8 shows the Nash-Sutcliffe Efficiency Coefficient, averaged for eight selected categories of streamflow, for the CONTROL and GEFS mean simulations. The CONTROL simulation has positive skill over all significant river basins for the first couple of days (upper panel). At day 7, the skill diminishes for all rivers except the three largest categories. With the ensemble mean as a deterministic forecast (lower panel), positive skill is associated with all significant river basins up to 16 days. The improvement of streamflow simulation due to the use of EPS, or the uncertainty in the initial conditions of the atmospheric model, is thus confirmed.

c. Probabilistic forecast skill CRPSS

CRPS and the corresponding skill score CRPSS are calculated for each grid point and each lead time. CRPSS is then averaged in space for each category of grid points and displayed in Fig. 9. Apparently, the score is dependent on the lead time and river size. The CONTROL simulation has some skill for all lead times for the largest river basins, but only for the first one to three days for all other rivers (upper panel). Comparison suggests that higher resolution GFS simulation is superior for the 3-7 days lead time over small and medium sized basins (not shown), and the mean of the GEFS member simulations has slightly higher skill than the CONTROL in the 2nd week (not shown). However, the major improvement is associated with the probabilistic forecast using the 10-member ensemble (lower panel).

From the definition of CRPS and CRPSS, lack of skill is partly due to the difference between the analysis and the simulation, or its mode in the case of ensemble based probabilistic forecast. Therefore, reduction in the systematic error of simulation can improve the CRPSS skills. For atmospheric variables, operational experience (e.g. Cui et al., 2005) and research with a synthetic data set (Son et al., 2008) suggest that various schemes can be used to achieve this goal. As most of these schemes require a longer time series as independent training data set, a simple method is used in this study. The bias is removed by subtracting the temporal mean of the simulation-analysis difference from each simulation time series. Using a dependent training date set, this procedure is not operationally applicable. Rather, it represents an upper limit of improvement and most importantly, provides a way to understand the deficit of the simulations.

Fig. 10 compares the CRPSS of the GEFS ensemble mean simulation before and after the bias-correction. To make it simple, only selected categories of grid points are plotted. For the raw simulation, only the three top categories, or the large river basins with mean streamflow of $500 \text{ m}^3 \text{ s}^{-1}$ and higher, have positive skill compared with the persistent forecast. Other categories are characterized by negative skills between day 2~4 to day 10~13. This medium-range deficit of the simulations can be overcome by bias correction. After removing the temporal mean error, positive skill over the persistent forecast is seen for all ranges of river sizes. However, the skills are still relatively low for lead time 3 to 7 days for the river basins with mean streamflow less than $500 \text{ m}^3 \text{ s}^{-1}$.

As the CRPS is also affected by the ensemble spread, the lack of skills for medium ($55\text{-}500 \text{ m}^3 \text{ s}^{-1}$) and small ($10\text{-}55 \text{ m}^3 \text{ s}^{-1}$) river basins at medium range (three to seven day) lead times may be due to insufficient spread of the simulated streamflow. An inspection of individual simulation trajectories of some medium and small rivers supports this hypothesis. For example, the May 4th GEFS simulation trajectories of the Potomac River at Washington DC are so close to each other (Fig. 3) that the analysis falls outside the range of the ensemble members from May 6th to 11th. This is even clearer with the Nehalem River at Foss, OR. Fig. 12 shows various simulated streamflow together with the analysis in

a manner similar to Fig. 3, but the 10 GEFS ensemble members are ranked at each lead time and each forecast case (initialization) before they are averaged over the 60 cases for each of the 10 ranks. By considering this ranked ensemble, the ensemble spread is largely conserved. The analysis falls outside the range of the ensemble members from day one through day eight. It should be pointed out that the horizontal resolution of the GEFS forecast is not high enough to predict the precipitation patterns and intensity at the scale comparable to a small river basin like the Nehalem River, and except for the largest river basins (mean streamflow over $500 \text{ m}^3 \text{ s}^{-1}$) downscaling is necessary when streamflow is to be predicted from the products of such low resolution models.

5. Summary and Discussions

In this article, the atmospheric component of a coupled ensemble forecasting system is evaluated in terms of its ability to provide reasonable forcing to the hydrological component and the effect of the uncertainty represented in the atmospheric ensemble system on the predictability of streamflow as a hydrological variable. The Global Ensemble Forecast System (GEFS) of NCEP is evaluated following a “perfect hydrology” approach, in which the hydrological model is considered free of errors and the initial conditions in the hydrological variables are assumed accurate, and the evaluation is performed over the Continental United States (CONUS) domain. From the results presented in section 4, the scientific questions listed in the introduction can be answered.

(1) Is the quality of the coupled system sufficient to generate any useful streamflow forecast? The GEFS ensemble forecasting system based on the coupled GFS-Noah model, with a river routing model attached, is able to reasonably simulate the analyzed streamflow for river basins with significant streamflow ($10 \text{ m}^3 \text{ s}^{-1}$ or above). Therefore, if any robust land surface model is coupled with the GFS model, and a good river routing model is attached to the coupled system, it is feasible to generate gridded streamflow forecast with usable skills.

(2) How is the skill of streamflow forecast dependent on the river basin size and lead time?

Generally speaking, the skill is higher for larger rivers and shorter lead times. For Small river basins (mean streamflow of $10\text{-}55\text{ m}^3\text{ s}^{-1}$) some useful skill can be achieved if the lead time is less than three days; useful forecast is possible with medium range lead time (three to seven days) for medium-sized river basins ($55\text{-}500\text{ m}^3\text{ s}^{-1}$) and predictability can be extended to extended range (the second week) for large river basins (above $500\text{ m}^3\text{ s}^{-1}$).

(3) Is the uncertainty represented in the ensemble generation strategy helpful to improve the streamflow forecast, and is it sufficient to account for the uncertainties in the hydrological output? The GEFS ensemble mean forecasts, and especially the GEFS ensemble based probabilistic forecasts, have more skill than single forecasts, indicating positive effect of the uncertainty represented in the atmospheric initial conditions on the streamflow forecast. However, it turns out the uncertainty in the atmospheric forcing is insufficient except for the large river basins.

(4) What measures should be taken in modifying the atmospheric modeling component to improve the ensemble streamflow forecasting? Systematic error is a significant part of the total forecast error, especially for medium and small river basins, and it can be reduced through a suitable bias-correction algorithm. For medium and small river basins, the medium range forecasts suffer from considerable under-dispersion, and improvement of the EPS strategy and downscaling of the atmospheric forcing, may lead to improvement.

The grouping of the grid points in to different categories of river basins based on the mean streamflow over the two month experimental period is somehow arbitrary and the inferred river basins size from these categories may be slightly different from the real catchment areas. Nevertheless, the smooth variation of verification scores with respect to the river size categories suggests the grouping is reasonable. In interpreting the results of the study, the river size can be viewed in the sense of both catchment area and the magnitude of streamflow. Therefore, the coupled system may have higher skills during the high flow periods than the low flow periods and in deed, case studies suggest the system has

skills in predicting extreme events. These questions are not fully studied in the article and may be topics for further investigations.

The evaluation of the GEFS ensemble forecast system is performed with a “perfect hydrology” approach and it addresses only the predictability of the atmospheric component of the coupled system, and thus the results are independent of the river routing model and the land surface model. Therefore, the conclusions of this research can be used as guidance in improving the atmospheric component of the coupled ensemble system for skilled hydrological forecast. A post processing scheme aimed at reducing systematic errors has been applied to meteorological variables other than precipitation in the North America Ensemble Forecast System (NAEFS) since 2006. It is an adaptive procedure applied at each grid point and thus automatically takes the spatial variation and weather/climate regimes into consideration (Cui et al., 2005). Similar scheme for precipitation and an algorithm for downscaling are under development at NCEP. There is also a plan to apply these procedures during model integration to benefit coupled meteorological-hydrological forecasts. A stochastic perturbation scheme will be implemented to the GEFS ensemble system and it will increase the ensemble spread (Hou et al., 2008) to better match the error of the ensemble mean and improve probabilistic forecast. With all these improvements in the atmospheric component, in addition to the developments in hydrological modeling and analysis, the coupled GFS-Noah ensemble forecast system in the near future is likely to be able to generate gridded streamflow forecast guidance with a probabilistic format.

Acknowledgements: The work on this research by NCEP/EMC was supported by the NOAA Climate Program Office funding for the NOAA Core Project for CPPA/GAPP (PI K. E. Mitchell) and the NOAA THORPEX program. We appreciate the discussions with D.-J. Seo, P. Restrepo, J. C. Schaake, A. Wood, D. Lettenmaier and Y. Xia, whose comments and suggestions helped to understand and interpret the results. Suggestions for improving the manuscript by Y. Xia and three anonymous peer

reviewers are acknowledged. The authors would like to thank Yuejian Zhu for help with data archives and George Gayno for assistance with the runoff interpolation.

REFERENCES

- Buizza, R., 1997: Potential forecast skill of ensemble prediction, and spread and skill distribution of ECMWF Ensemble prediction system. *Mon. Wea. Rev.*, **125**, 99-119.
- Buizza, R., M. J. Miller and T. Palmer, 1999: Stochastic simulation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. Meteorol. Soc.*, **125**, 2887-2908.
- Cui, B., Z. Toth, Y. Zhu, D. Hou, and S. Beauregard, 2005: Statistical post-processing of operational and CDC hindcast ensembles, Preprints, 21st Conference on weather Analysis and Forecasting/17th Conference on Numerical Weather Prediction. Washington DC, Aug 1-5, 2005, *Amer. Meteor. Soc.*, **12B.2**, 2005.
- Gourley, J. J. and Vieux B. E., 2005: A method for evaluating the accuracy of quantitative precipitation estimates from a hydrologic perspective. *J Hydrometeorol.*, **6**, 115-133.
- Hou, Dingchen, Zoltan Toth, Yuejian Zhu and Weiyu Yang, 2008: Impact of a Stochastic Perturbation Scheme on NCEP's Global Ensemble Forecast System. In Preprints, *AMS 88th Annual Meeting and 19st AMS conference on Probability and Statistics*, 20-24 January, 2008, New Orleans, LA.
- Houtekamer, P. L., L. Lefavre, J. Derome, H. Ritchie, and H.L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225-1242.
- Houtekamer, P. L. and H.L. Mitchell, 1997: Using ensemble forecasts for model validation. *Mon. Wea. Rev.*, **125**, 3-26.
- Koutsoyiannis, D., A. Efstratiadis and K. Georgakakos, 2007: Uncertainty assessment of future hydroclimatic predictions: A comparison of probabilistic and scenario-based approaches. *J Hydrometeorol.*, **8**, 261-281.
- Krzysztofowicz R, 2002: Bayesian system for probabilistic river stage forecasting. *J Hydrol*, **268**, 16-40.
- Lohmann, D. K. E. Mitchell, P. R. Houser, E. F. Wood, J. C. Schaake, A. Robock, B. A. Cosgrove, J. Sheffield, Q. Duan, L. Luo, R. W. Higgins, R. T. Pinker, J. D. Tarpley, 2004: Streamflow and

- water balance intercomparison of four land surface models in the North American Land Data Assimilation Systems project. *J. Geophys. Res.*, **109**, D07S91, doi:10.1029/2003JD003517.
- Milly, P. C. D., and K. A. Dunne, 2002a: Macroscale water fluxes 1. Quantifying errors in the estimation of basin mean precipitation. *Water Resources Research*, **38**, No 10, 1205 doi:10.1029/2001WR000759.
- Milly, P. C. D., and K. A. Dunne, 2002b: Macroscale water fluxes 2. Water and energy supply control of their interannual variability. *Water Resources Research*, **38**, No.10, 1206 doi:10.1029/2001WR000760.
- Mitchell, K. E., D. Lohmann, P. R. Houser, E. F. Wood, J. C. Schaake, A. Robock, B. A. Cosgrove, J. Sheffield, Q. Duan, L. Luo, R. W. Higgins, R. T. Pinker, J. D. Tarpley, D. P. Lettenmaier, C. H. Marshall, J. K. Entin, M. Pan, W. Shi, V. Koren, J. Meng, B. H. Ramsay, and A. A. Bailey, 2004: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCM products and partners in a continental distributed hydrological modelling system. *J. Geophys. Res.*, **109**, D07S90, doi:10.1029/2003JD003823.
- Mitchell, K. E., H. Wei, S. Lu, G. Gayno and J. Meng, 2005: NCEP implements major upgrade to its medium-range global forecast system, including land-surface component. *GEWEX newsletter*, May 2005.
- Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, T. L. Veith, 2007: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, **50**, 885-900.
- Nash, J. E. and J. V. Sutcliffe, 1970: River flow forecasting through conceptual models, Part I: A discussion of principles, *Journal of Hydrology*, **10**, 282-290.
- Pappenberger, F., J. Bartholmes, J. Thielen, H. L. Cloke, R. Buizza, and A. de Roo, 2008: New dimensions in early warning across the globe using grand-ensemble weather predictions. *Geophys. Res. Lett.*, **35**, L10404, doi:10.1029/2008GL033837.

- Pappenberger F, Beven K. J., Hunter N, P. D. Bates, B. T. Couweleew, J. Thielen and A. P. J. de Roo, 2005: Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrology and Earth System Science*, **9**, 381-393.
- Son, J., D. Hou and Z. Toth, 2008: Evaluation of a Bayesian Bias Estimator. Submitted to *Nonlinear Processes in Geosciences*.
- Toth, Z. and E. Kalnay, 1997: Ensemble forecasting at NCEP: The Breeding Method. *Mon. Wea. Rev.*, **125**, 3297-3318.
- Toth, Z., O. Talagrand, G. Candille and Y. Zhu, 2003: Probability and Ensemble Forecast. *Forecast Verification: A practitioner's Guide in Atmospheric Science*. Jolliffe, I. T. and D. B. Stephenson Ed., John Wiley & Sons Ltd.
- Verbunt, M., M. Zappa, J. Gurtz, P. Kaufmann, 2006: Verification of a coupled hydrometeorological modelling approach for Alpine tributaries in the Rhine basin. *J. Hydrol.*, **324**, 224-238.
- Werner, K., D. Brandon, M. Clark and S. Gangopadhyay, 2005: Incorporating medium-range numerical weather model output into the ensemble streamflow prediction system of the National Weather Service. *J Hydrometeorol.*, **6**, 101-114.

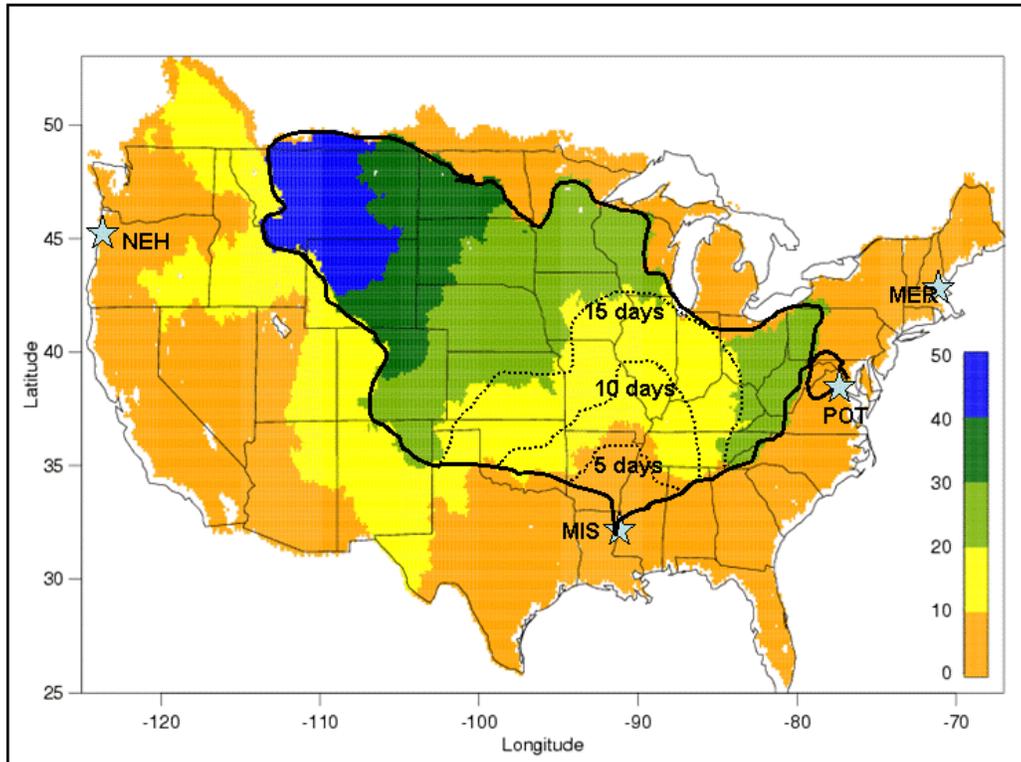


Fig. 1. Map showing travel time to outlet (to Oceans or the Great Lakes) with information for the river basins used as examples. The shading is the travel time of surface water from each grid cell to its downstream outlet (to oceans and great lakes) in unit of days (from Lohmann et al., 2004). The stars mark the approximate locations of the grid cells, representing river basins MIS, MER, POT and NEH, listed in Table 2. The upstream catchments of the rivers MIS and POT are approximately outlined by the heavy solid curves, while those for MER and NEH are too small to be shown. Inside the MIS catchment, the travel time to the MIS grid cell is schematically shown with estimated contours of 5, 10 and 15 days (dashed contours).

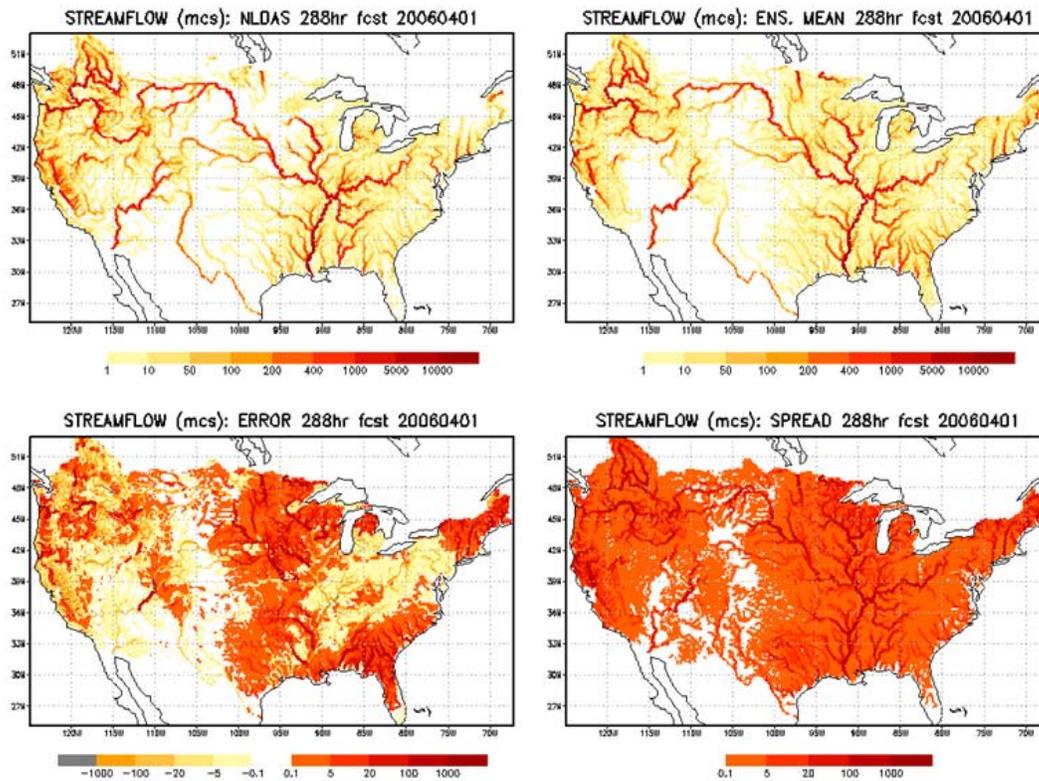


Fig. 2. Streamflow (m^3s^{-1}) according to analysis (upper left) and the ensemble mean of GEFS simulations (upper right), their difference (simulation-analysis, lower left) and GEFS ensemble spread (lower right). Note that the shading scale in the lower panels is different than that in the upper panels. The simulations are 288h prognosis initialized at 0000 UTC, 1 April, 2006, and the analysis is valid at the same time as the simulations (0000 UTC, 13 April, 2006).

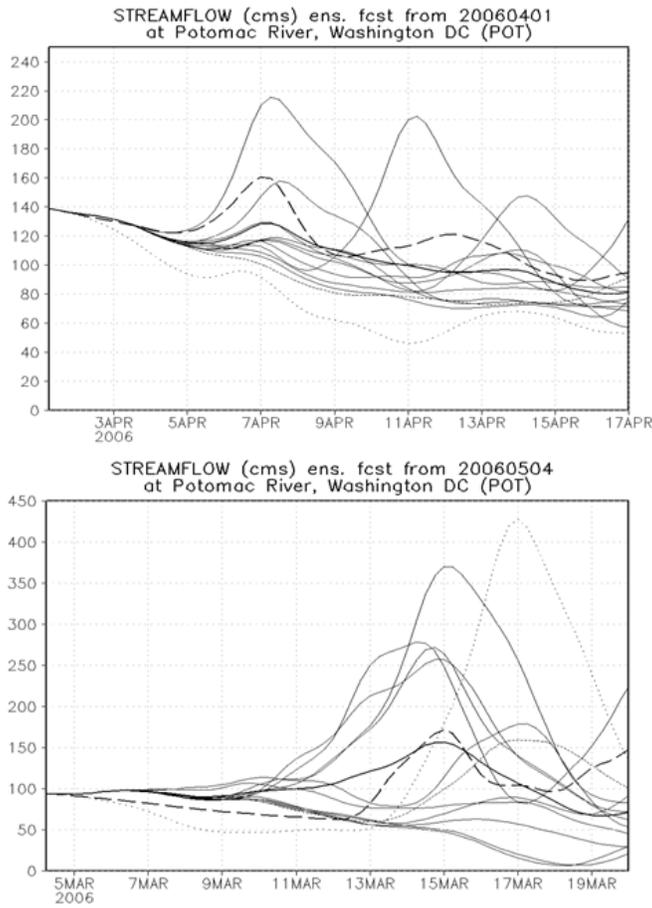


Fig. 3. Simulated streamflow initialized at 0000 UTC, 1 April (upper panel) and 4 May (lower panel), 2006, at the Potomac River, Washington DC, are shown as a function of time. The thin curves are for single simulations GFS (dotted), CONTROL (dashed) and GEFS members (solid). The heavy solid curve is the GEFS ensemble mean and the analyzed streamflow is shown as the heavy dashed curve.

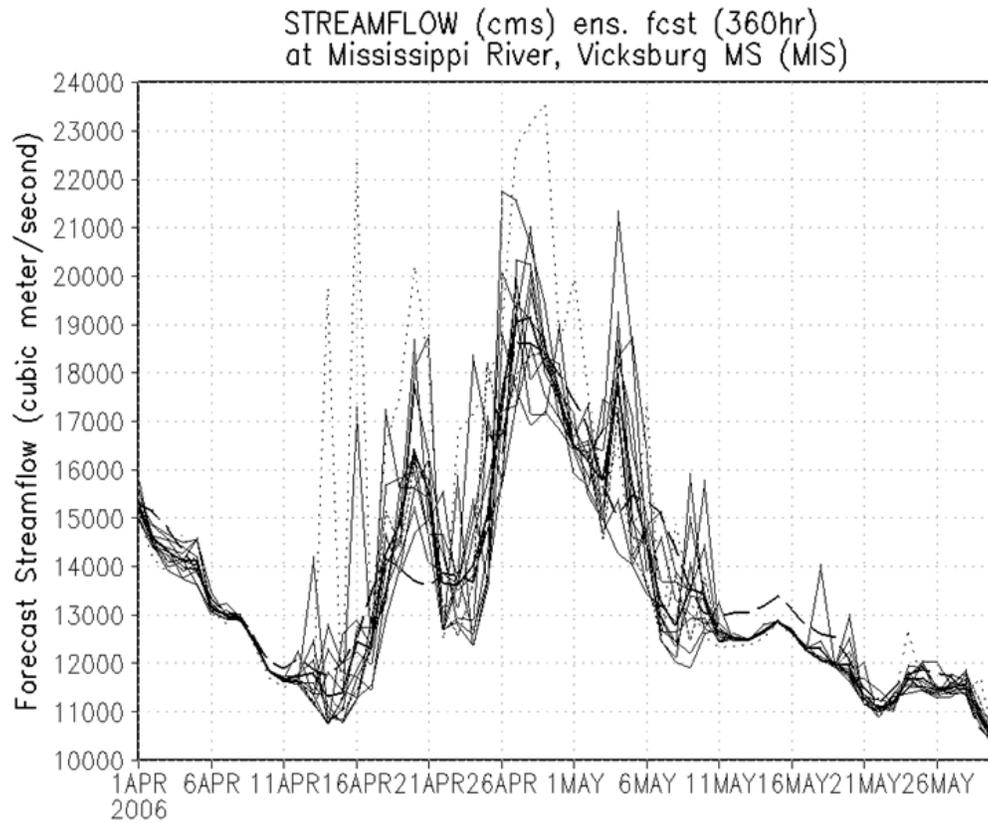


Fig. 4. Daily values of simulated streamflow for lead time of 15 days, for the grid cell at the lower Mississippi River, Vicksburg, MS, with the corresponding analysis. The thin curves are for single simulations GFS (dotted), CONTROL (dashed) and GEFS members (solid). The heavy solid curve is the GEFS ensemble mean and the analyzed streamflow is shown as the heavy dashed curve.

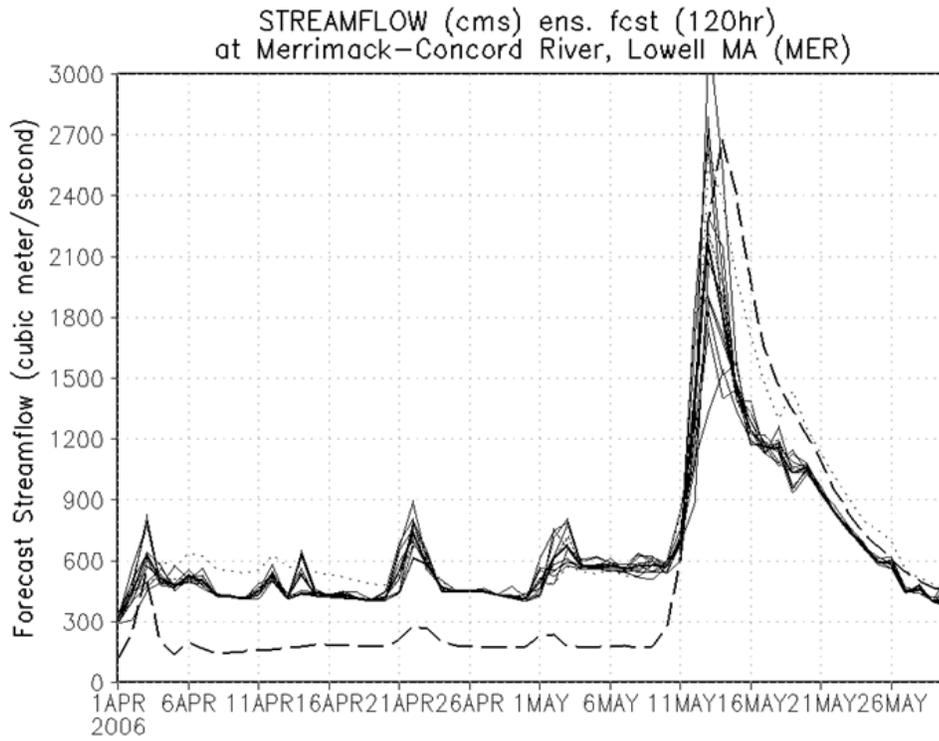


Fig. 5. Same as Fig. 4, except for 5-day simulations for the grid cell corresponding to the Merrimack-Concord River at Lowell, MA.

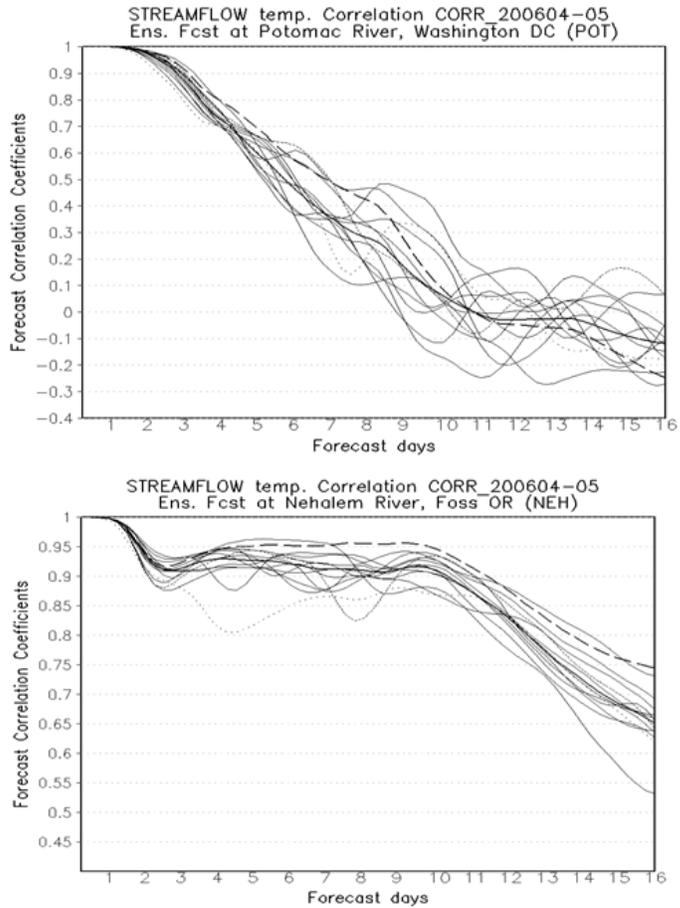


Fig. 6. Correlation coefficients between simulated and analyzed streamflow, as a function of lead time, at Potomac River, Washington DC (upper panel) and Nehalem River, Foss, OR (lower panel). The thin curves are for single simulations GFS (dotted), CONTROL (dashed) and GEFS members (solid). The heavy dashed curve is for the GEFS ensemble mean and the heavy solid curve depicts the average scores of individual GEFS members.

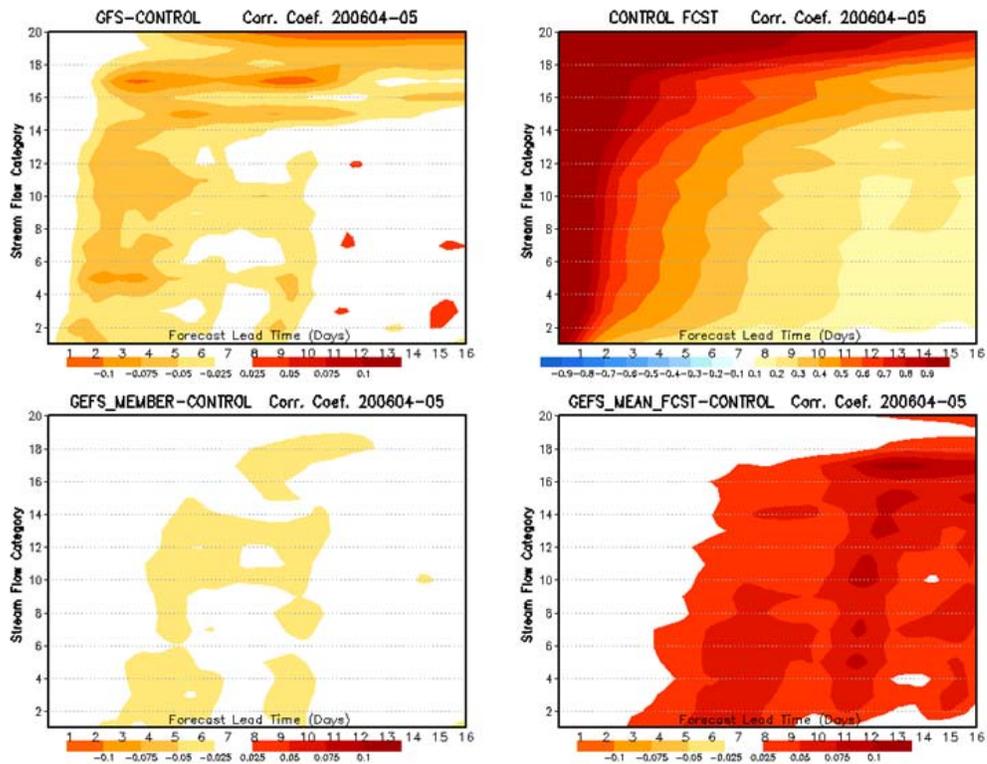


Fig. 7. The upper right panel shows the correlation coefficients of the CONTROL simulation of streamflow, as a function of lead time and streamflow category. Other panels show the deviation from this score, by using GFS (upper left), the mean score of the 10 GEFS members (lower left) and the GEFS ensemble mean simulation (lower right).

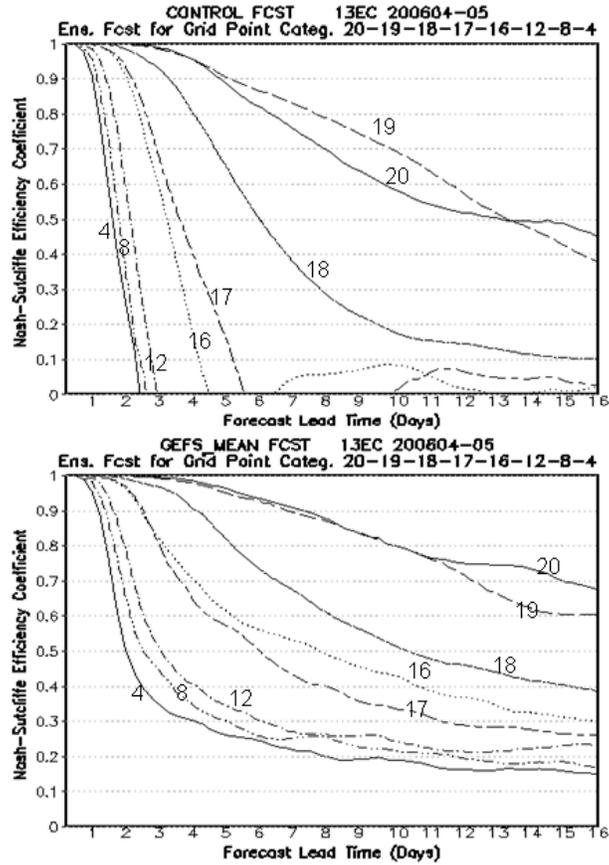


Fig. 8. Nash-Sutcliffe Efficiency Coefficient, averaged over selected river size categories, of the CONTROL (upper panel) and the GEFS mean (lower panel) streamflow, as function of lead time. The category index is marked on the corresponding curve.

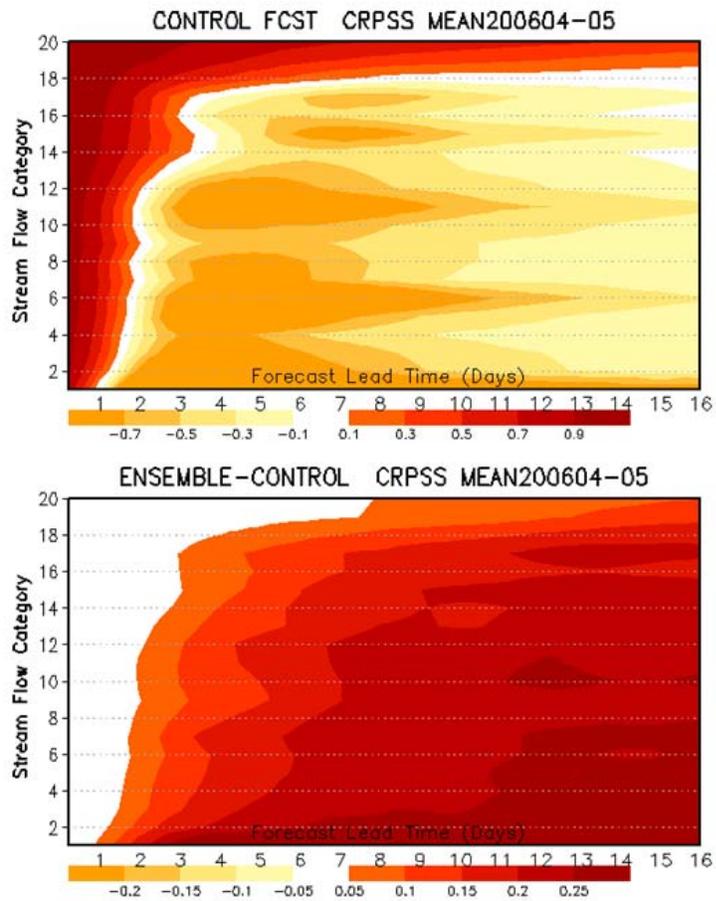


Fig. 9. CRPSS of the CONTROL simulation of streamflow, as function of lead time and river size category (upper panel) and the deviation from this score by using the GEFS ensemble based probabilistic forecast.

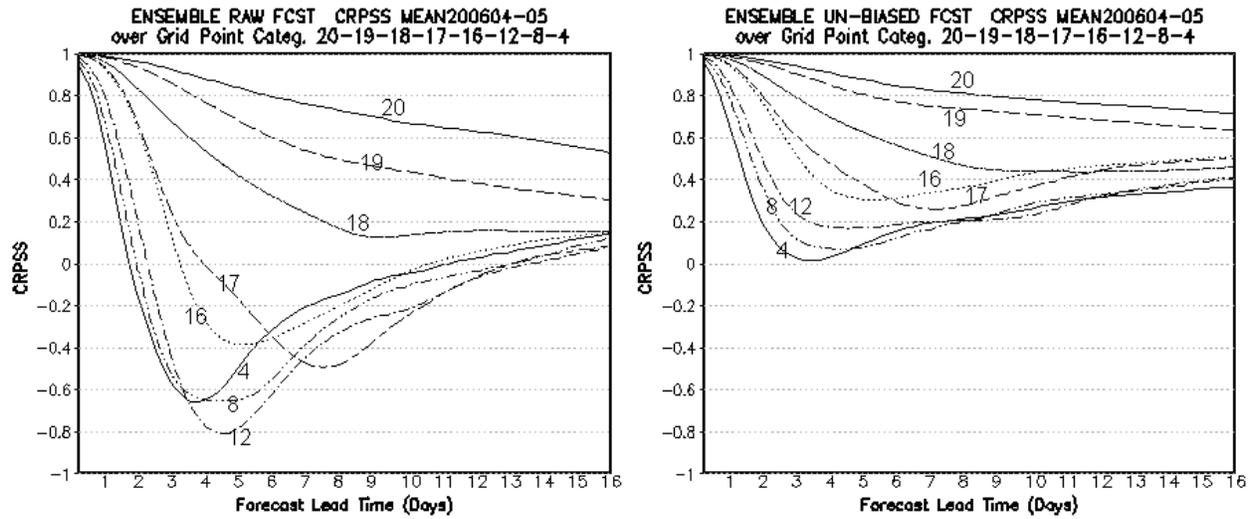


Fig. 10 CRPSS averaged over selected ranges of mean streamflow, calculated from the raw (left) and bias-removed (right) GEFS 10-member ensemble forecasts, as function of lead time. The category index is marked on the corresponding curve.

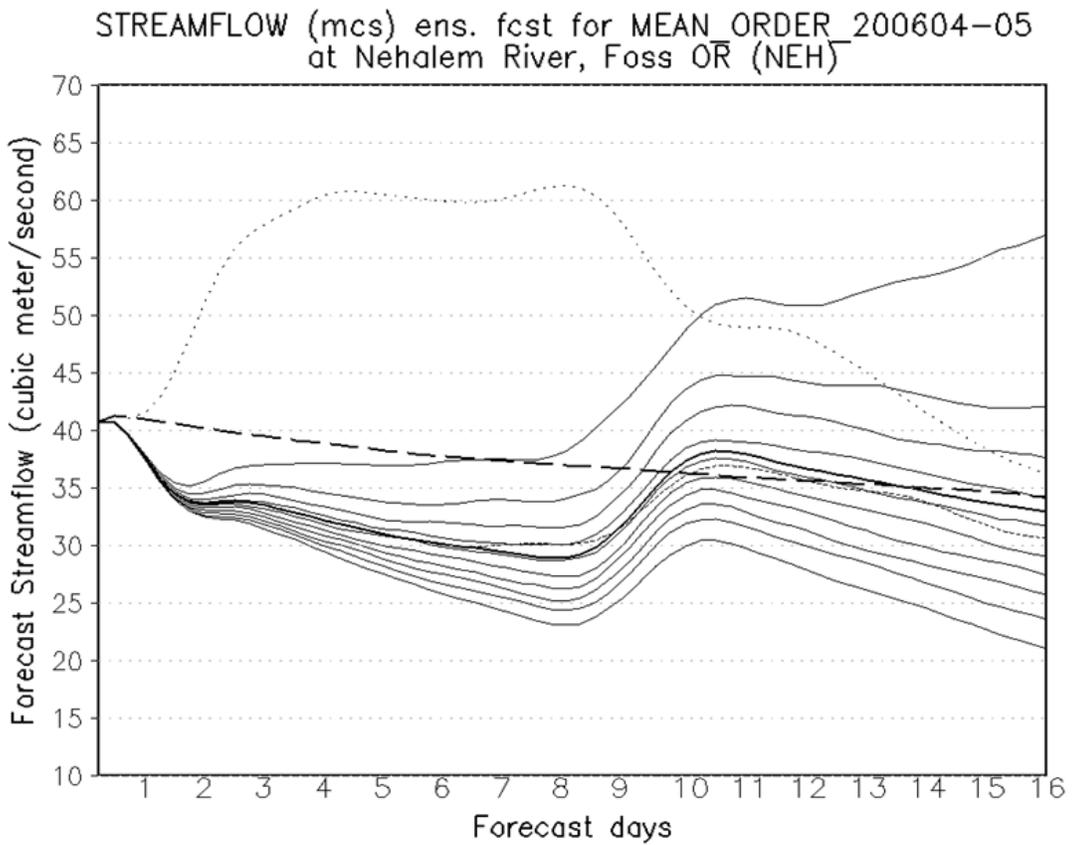


Fig.11. Simulated streamflow averaged over all 60 cases at the Nehalem Rive at Foss, as a function of lead time. The thin curves are for single simulations GFS (dotted), CONTROL (dashed) and GEFS members (solid). The heavy solid curve is the GEFS ensemble mean and the analyzed streamflow is shown as the heavy dashed curve.

Table 1. The 20 categories of grid points, grouped based on the analysed streamflow averaged over the period of experiment.

| Category | Range of Mean Streamflow ($\text{m}^3 \text{s}^{-1}$) | Number of Grid Points | Percentage of Grid Points |
|----------|---|-----------------------|---------------------------|
| 1 | 0.0 ~ 1.0 I | 71473 | 68.77 |
| 2 | 1.0 ~ 10.0 | 18481 | 17.78 |
| 3 | 10.0 ~ 15.0 | 2454 | 2.36 |
| 4 | 15.0 ~ 20.0 | 1558 | 1.50 |
| 5 | 20.0 ~ 25.0 | 985 | 0.95 |
| 6 | 25.0 ~ 30.0 | 800 | 0.77 |
| 7 | 30.0 ~ 35.0 | 621 | 0.60 |
| 8 | 35.0 ~ 40.0 | 506 | 0.49 |
| 9 | 40.0 ~ 45.0 | 433 | 0.42 |
| 10 | 45.0 ~ 55.0 | 723 | 0.70 |
| 11 | 55.0 ~ 70.0 | 665 | 0.64 |
| 12 | 70.0 ~ 90.0 | 667 | 0.64 |
| 13 | 90.0 ~ 120.0 | 593 | 0.57 |
| 14 | 120. ~ 150.0 | 533 | 0.51 |
| 15 | 150.0 ~200.0 | 634 | 0.61 |
| 16 | 200.0 ~ 300.0 | 681 | 0.66 |
| 17 | 300.0 ~ 500.0 | 592 | 0.57 |
| 18 | 500.0 ~ 1000.0 | 557 | 0.54 |
| 19 | 1000.0 ~ 2000.0 | 517 | 0.50 |
| 20 | 2000.0 and up | 463 | 0.45 |

Table 2. Information of grid cells used in the text as examples of river basins

| Abbr. | Name of USGS River Station | Mean Streamflow ($\text{m}^3 \text{s}^{-1}$) | Category |
|-------|------------------------------------|--|----------|
| NEH | Nehalem River, FOSS OR | 40.3 | 9 |
| POT | Potomac River, Washington DC | 110.7 | 13 |
| MER | Merrimack-Concord River, Lowell MA | 499.4 | 17 |
| MIS | Mississippi River, Vicksburg MS | 14788.3 | 20 |

Figure Captions

Fig. 1. Map showing travel time to outlet (to Oceans or the Great Lakes) with information for the river basins used as examples. The shading is the travel time of surface water from each grid cell to its downstream outlet (to oceans and great lakes) in unit of days (from Lohmann et al., 2004). The stars mark the approximate locations of the grid cells, representing river basins MIS, MER, POT and NEH, listed in Table 2. The upstream catchments of the rivers MIS and POT are approximately outlined by the heavy solid curves, while those for MER and NEH are too small to be shown. Inside the MIS catchment, the travel time to the MIS grid cell is schematically shown with estimated contours of 5, 10 and 15 days (dashed curves).

Fig. 2. Streamflow (m^3s^{-1}) according to analysis (upper left) and the ensemble mean of GEFS simulations (upper right), their difference (simulation-analysis, lower left) and GEFS ensemble spread (lower right). Note that the shading scale in the lower panels is different than that in the upper panels. The simulations are 288h prognosis initialized at 0000 UTC, 1 April, 2006, and the analysis is valid at the same time as the simulations (0000 UTC, 13 April, 2006).

Fig. 3. Simulated streamflow simulation initialized at 0000 UTC, 1 April (upper panel) and 4 May (lower panel), 2006, at the Potomac River, Washington DC, are shown as a function of time. The thin curves are for single simulations GFS (dotted), CONTROL (dashed) and GEFS members (solid). The heavy solid curve is the GEFS ensemble mean and the analyzed streamflow is shown as the heavy dashed curve.

Fig. 4. Daily values of simulated streamflow for lead time of 15 days, for the grid cell at the lower Mississippi River, Vicksburg, MS, with the corresponding analysis. The thin curves are for single simulations GFS (dotted), CONTROL (dashed) and GEFS members (solid). The heavy solid curve is the GEFS ensemble mean and the analyzed streamflow is shown as the heavy dashed curve.

Fig. 5. Same as Fig. 4, except for 5-day simulations for the grid cell corresponding to the Merrimack-Concord River at Lowell, MA.

Fig. 6. Correlation coefficients between simulated and analyzed streamflow, as a function of lead time, at Potomac River, Washington DC (upper panel) and Nehalem River, Foss, OR (lower panel). The thin curves are for single simulations GFS (dotted), CONTROL (dashed) and GEFS members (solid). The heavy dashed curve is for the GEFS ensemble mean and the heavy solid curve depicts the average scores of individual GEFS members.

Fig. 7. The upper right panel shows the correlation coefficients of the CONTROL simulation of streamflow, as a function of lead time and streamflow category. Other panels show the deviation from this score, by using GFS (upper left), the mean score of the 10 GEFS members (lower left) and the GEFS ensemble mean simulation (lower right).

Fig. 8. Nash-Sutcliffe Efficiency Coefficient, averaged over selected river size categories, of the CONTROL (upper panel) and the GEFS mean (lower panel) streamflow, as function of lead time. The category index is marked on the corresponding curve.

Fig. 9. CRPSS of the CONTROL simulation of streamflow, as function of lead time and river size category (upper panel) and the deviation from this score by using the GEFS ensemble based probabilistic forecast.

Fig. 10 CRPSS averaged over selected ranges of mean streamflow, calculated from the raw (left) and bias-removed (right) GEFS 10-member ensemble forecasts, as function of lead time. The category index is marked on the corresponding curve.

Fig.11. Simulated streamflow averaged over all 60 cases at the Nehalem Rive at Foss, as a function of lead time. The thin curves are for single simulations GFS (dotted), CONTROL (dashed) and GEFS members (solid). The heavy solid curve is the GEFS ensemble mean and the analyzed streamflow is shown as the heavy dashed curve.