NOAA'S SECOND-GENERATION GLOBAL MEDIUM-RANGE ENSEMBLE REFORECAST DATASET

BY THOMAS M. HAMILL, GARY T. BATES, JEFFREY S. WHITAKER, DONALD R. MURRAY, MICHAEL FIORINO, THOMAS J. GALARNEAU JR., YUEJIAN ZHU, AND WILLIAM LAPENTA

A new set of NOAA reforecasts—now featuring a current operational model and a wider set of variables archived in higher resolution—is freely accessible to the weather forecast community.

"Those who cannot remember the past are condemned to repeat it."

—George Santayana

The weather and climate prediction community have made continued, significant improvement in the quality of numerical forecast guidance. This has come as a result of increased resolution; improved physical parameterizations; improved chemistry and aerosol physics; improved estimates of the initial state estimate due to better data assimilation techniques; and improved couplings between the atmosphere and the land surface, cryosphere, ocean, and more. Nonetheless, judging from the pace of past improvements, medium-range forecast systematic errors will not become negligibly small within the next decade or two. For intermediate-resolution simulations such as those from current-generation global ensemble systems, users of forecast guidance may notice biased surface temperature forecasts, precipitation forecasts with insufficient detail in mountainous terrain, or perhaps too much drizzle or too little heavy rain. They may notice over- or underestimated cloud cover or that near-surface winds are characteristically much stronger than forecast. They may notice that hurricanes are too large in size but less intense than observed. Sometimes, however, systematic errors may be less obvious. Does the model forecast of the Madden–Julian oscillation (MJO; Zhang 2005) propagate too slowly or decay too quickly? Are Arctic cold outbreaks too intense, and do they plunge south too quickly or too slowly? Does the model overforecast the frequency of tropical cyclogenesis in the Caribbean Sea? Do tropical cyclones tend to recurve too quickly or slowly? Such questions may be difficult to answer quantitatively with a month or even a year of model guidance.

In such circumstances, reforecasts can be used to great advantage to distinguish between the random and the model errors. Reforecasts are especially helpful for statistically adjusting weather and climate forecasts to observed data, ameliorating the errors and improving objective guidance (Hamill et al. 2006; Hagedorn 2008). Reforecasts, also commonly called hindcasts, are retrospective forecasts for many dates in the past, ideally conducted using the same forecast model and same assimilation system used operationally.[1] Reforecasts have been shown to be

[1] We prefer the term "reforecast" in this instance to "hindcast" so as to make the association in the reader's mind with reanalyses. This reforecast would not have been very useful were there not a high-quality reanalysis to provide initial conditions, here from the NCEP Climate Forecast System Reanalysis.

particularly useful for the calibration of relatively uncommon phenomena such as heavy precipitation (Hamill et al. 2008) and longer-lead weather–climate phenomena (Hamill et al. 2004), where there is small forecast signal and comparatively large noise owing to chaos and model error. In both cases, the large sample size afforded by reforecasts is useful for finding a suitably large number of past similar forecast scenarios. With associated observational data, one then can estimate a conditional distribution of the possible observed states given today's numerical guidance, assuming past forecasts have similar errors to current forecasts. Even when no observed data are available for calibration, reforecasts can be useful for determining the climatology of a model. A 20 m s^{-1} surface wind would be exceptionally strong in most locations on Earth, but if the forecast model severely overforecasts wind speeds, such an event may be of less concern. A reforecast can thus be used for estimating the forecast climatology, placing the current forecast in context (Lalaurette 2003, 2013).

The reforecast dataset discussed here makes an unprecedentedly large volume of data accessible to users. Over 27 years of once-daily, 11-member ensemble forecasts were computed using the same model version, the same uncertainty parameterization, and a very similar method of ensemble initialization to the currently operational National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS). More than 125 TB of forecast output is conveniently available for fast-access download, and the full model dataset (~1 PB) is archived on tape. This dataset is more extensive than contemporary alternatives, such as the 5-member, ~20-yr weekly reforecasts from the European Centre for Medium Range Weather Forecasts (ECMWF; Hagedorn 2008; Hagedorn et al. 2012), and there is no charge for its use. Daily lagged reforecasts were also generated for the NCEP Climate Forecast System (CFS) seasonal forecasts (Saha et al. 2010).

We had several rationales for creating this extensive a reforecast dataset. The first is that we hope that the greater number of forecast samples from a statistically consistent model will lead to the diagnosis of model errors and development of novel and improved statistical calibration algorithms and algorithms for rare events and for novel applications—algorithms that may be less accurate were they developed with smaller training datasets. An example of this is products for the renewable energy sector, such as extended-range wind and solar energy potential forecasts. We also hope that by making these data and experimental products from it freely available, the dataset will be used widely.

A second major reason for generating this extensive dataset was to quantify the benefits of this additional training data. Do we really need an exceptionally large training sample size, or might the products be acceptably similar in skill were they developed with a smaller reforecast dataset, perhaps with fewer members, fewer past years, or skipping days between samples? Generating a large reforecast dataset is computationally expensive and labor intensive. For this dataset, more than 15 million CPU hours were used on the Department of Energy's Lawrence Berkeley Laboratory supercomputers, and approximately 5 person years of effort were expended to generate the reforecasts and set up the archives. Such extensive data may also not be an unalloyed benefit; the reforecasts in the distant past may have larger errors owing to a thinner observing network. Hence, should reforecasting become a regular component of National Weather Service's suite of numerical guidance, it will be helpful to determine the optimal configuration to apply to future ensemble forecast systems—the compromise that provides adequate training data to the statistical applications while being as computationally inexpensive as possible.

The next section of the article will discuss the contents of the dataset and the procedures to follow in order to download these data. We will then demonstrate some statistical characteristics of the raw reforecast dataset. The penultimate section describes several forecast applications. The final section provides conclusions.

AFFILIATIONS: Hamill and Whitaker— Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado; Bates and Murray—Cooperative Institute for Research in the Environmental Sciences, University of Colorado, Boulder, Colorado; Fiorino—Global Systems Division, NOAA/Earth System Research Laboratory, Boulder, Colorado; Galarneau—National Center for Atmospheric Research,* Boulder, Colorado; Zhu and Lapenta—Environmental Modeling Center, NOAA/National Centers for Environmental Prediction, College Park, Maryland
* The National Center for Atmospheric Research is sponsored by the National Science Foundation.
CORRESPONDING AUTHOR: Dr. Thomas M. Hamill, NOAA Earth System Research Lab, Physical Sciences Division, R/PSD 1, 325 Broadway, Boulder, CO 80305
E-mail: tom.hamill@noaa.gov

The abstract for this article can be found in this issue, following the table of contents.
DOI:10.1175/BAMS-D-12-00014.1

In final form 1 February 2013

A DESCRIPTION OF THE REFORECAST DATASET AND HOW TO ACCESS IT.

The operational configuration of the NCEP GEFS changed as of 1200 UTC 14 February 2012. The real-time and reforecast models use version 9.0.1 of the GEFS, discussed at www.emc.ncep.noaa.gov/GFS/impl.php. For more detail on the GEFS, see Hamill et al. (2011a). During the first 8 days of the operational GEFS forecast and the reforecast, the model is run at T254L42 resolution, which with a quadratic Gaussian transform grid is an equivalent grid spacing of approximately 40 km at 40° latitude, and 42 vertical levels. Starting at day +7.5, the forecasts are integrated at T190L42, or approximately 54 km at 40° latitude, and data are saved at this resolution from days +8 to days +16—the end of the GEFS integration period. Note that there is a bug in version 9.0.1, resulting in the use of incorrect land surface tables in the land surface parameterization, which has introduced significant biases to near-surface temperatures. These errors are at least consistent between the current operational GEFS and the reforecast.

Through 20 February 2011, control initial conditions were generated by the Climate Forecast System Reanalysis (CFSR) (Saha et al. 2010). This used the Grid-Point Statistical Interpolation (GSI) system of Kleist et al. (2009) at T382L64. From 20 February 2011 through May 2012, initial conditions were taken from the operational GSI analysis, internally computed at T574L64. After 22 May 2012, the GSI was upgraded to use a hybrid ensemble Kalman filter–variational analysis system (Hamill et al. 2011a,b). This analysis improved the skill of operational GEFS forecasts and thus of the reforecasts introduced into the archive subsequent to that date.

The perturbed initial conditions for both the operational GEFS and the reforecast use the ensemble transform technique with rescaling (ETR) (Wei et al. 2008). For the operational real-time forecasts, 80 members are cycled for purposes of generating the initial condition perturbations. However, only the leading 20 perturbations plus the control initial condition were used to initialize the operational medium-range forecasts. The operational medium-range GEFS forecasts are generated every 6 h from 0000, 0600, 1200, and 1800 UTC initial conditions. In comparison, the reforecast was generated only once daily, at 0000 UTC, and only 10 perturbed forecast members and the one control forecast were generated. However, the 6-hourly cycling of ETR perturbations was preserved, though this cycling used only the 10 perturbed members rather than the 80 used in real time. Model uncertainty in the GEFS is estimated with the stochastic tendencies following Hou et al. (2008) for both operations and reforecasts.

Here are some details on the reforecast data that are available. About 29 years (December 1984–present) of reforecast data are currently archived. The archive includes the 0000 UTC GEFS real-time forecasts, which will be available with some delay, perhaps by 1300 UTC, though many fields will be available more quickly via the National Oceanic and Atmospheric Administration (NOAA)/National Operational Model Archive and Distribution System (NOMADS; http://nomads.ncdc.noaa.gov). Ninety-eight different forecast global fields are available at 1° resolution, and 28 selected fields are also available at the native resolution (~0.5° Gaussian grid spacing for the first week's forecasts and ~0.67° grid spacing for the second week's forecasts). Data are internally archived in GRIB2 format (www.nco.ncep.noaa.gov/pmb/docs/grib2/). The 1° data were created from the native-resolution data via bilinear interpolation using wgrib2 software (www.cpc.ncep.noaa.gov/products/wesley/wgrib2/). The listing of the fields that were saved and their resolutions are provided in Tables 1 and 2. Reforecast data were saved at 3-hourly intervals from 0 to 72 h and every 6 h thereafter. The 28+ years of data daily currently archived totals approximately 125 TB of internal storage.

Reforecast data can be accessed in many different ways. For users who want a few select fields (e.g., precipitation forecasts) spanning many days, months, or years, we provide a web interface for accessing such data (http://esrl.noaa.gov/psd/forecasts/reforecast2/). The interface allows the user to select particular fields, date ranges, domains, and type of ensemble information (particular members, the mean, or the spread). While data are internally archived in GRIB2 format, the synthesized files produced from a user's web form input are in netCDF format (www.unidata.ucar.edu/software/netcdf/). Should a user desire GRIB2 data instead, the raw data can be accessed via anonymous ftp (at ftp://ftp.cdc.noaa.gov/Projects/Reforecast2) or using wgrib2's "fast downloading" capabilities (www.cpc.ncep.noaa.gov/products/wesley/fast_downloading_grib.html). We request that users be conservative with their downloads in order to minimize computations and bandwidth.

Some users may desire only selected days of reforecasts but want full model output rather than the limited set of fields and levels available from Earth System Research Laboratory (ESRL). In this case, the user can download these data from the tape archive at the U.S. Department of Energy (the web form

Table 1. Reforecast variables available for selected mandatory and other vertical levels. Geopotential height is indicated by F, and an X indicates that this variable is available from the reforecast dataset at 1° resolution; a Y indicates that the variable is available at the native ~0.5° resolution. AGL indicates "above ground level." Hybrid sigma-pressure vertical levels (a very close approximation to sigma levels near the ground) are called "hyb."

Vertical level	U	V	T	F	q	Wind power
10 hPa	X	X	X	X		
50 hPa	X	X	X	X		
100 hPa	X	X	X	X		
200 hPa	X	X	X	X		
250 hPa	X	X	X	X		
300 hPa	X	X	X	X	X	
500 hPa	X	X	X	X	X	
700 hPa	X	X	X	X	X	
850 hPa	X	X	X	X	X	
925 hPa	X	X	X	X	X	
1000 hPa	X	X	X	X	X	
Hyb 0.996	X	X		X		
Hyb 0.987	X	X		X		
Hyb 0.977	X	X		X		
Hyb 0.965	X	X		X		
80 m AGL	X,Y	X,Y				X,Y

for this is at http://portal.nersc.gov/project/refcst/v2/). Such full data may be useful for, say, initializing high-resolution regional reforecasts. An example of this will be provided in the forecast applications section.

CHARACTERISTICS OF THE RAW REFORECAST DATA. The skill of the second-generation global ensemble reforecasts has improved very significantly from those from the first generation. Figure 1 shows a time series of yearly averaged global 500-hPa geopotential height anomaly correlations (AC) from both systems. For recent years, the day +5 second-generation reforecasts are more accurate than the day +3 first-generation reforecasts. Considering the second-generation reforecast, there is a modest change in average skill of the reforecasts during the 26-yr period shown. Yearly average AC increases in the version 2 reforecasts during the period with the change somewhat less than 1 day. For example, the day +5 forecasts for 2009/10 appear to be roughly comparable to the day +4 forecasts (not shown) from 1985 to 1986. This is likely due primarily to changes in the observing network and observation data processing during the reanalysis period (Wang et al. 2011; Kumar et al. 2012).

Tropical cyclone forecast tracks were calculated using the GFDL tracker algorithm (Gopalakrishnan et al. 2012). Figure 2 shows track statistics binned by half decades. There has been a pronounced improvement in track forecasting during the period of the reforecasts. This is at least in part due to greater changes in the forecast skill of the steering flow in the tropics, owing to improvements in the CFSR analyses over time. Tropical 500-hPa geopotential height anomaly correlations improved by 1–2 days between 1985/86 and 2009/10 (not shown). Such large changes in skill during the reforecast period can make it more difficult to achieve high forecast accuracy with simple statistical postprocessing algorithms, for the forecast errors in past cases will not be fully representative of current forecast errors. Some of these differences, however, also might be due to a change in the accuracy of the observed locations; past observed tracks may not be as accurate as more recent observed tracks. Our own internal computations of blended climatology and persistence track forecasts (CLIPER; Neumann 1972) shows that western Pacific CLIPER track errors have also decreased substantially in the past 25 years.

REFORECAST APPLICATIONS. We anticipate that many groups will use this reforecast dataset to explore, compare, and validate methods for statistically postprocessing the model data. Here we consider the usage of the reforecast for postprocessing 24-h accumulated precipitation forecasts, both probabilistic and deterministic.

Previously, an analog technique was demonstrated with the first-generation reforecasts as one of many possible methods for statistically downscaling and correcting the forecasts, improving their reliability and skill (Hamill et al. 2006; Hamill and Whitaker 2006). Figure 3 shows Brier skill scores from the first- and second-generation reforecasts, processed using the rank analog technique described more

Table 2. Single-level reforecast variables archived (and their units). Where a [Y] is displayed, this indicates that this variable is available at the native ~0.5° resolution as well as the 1° resolution.
Variable (units)
Mean sea level pressure (Pa) [Y]
Skin temperature (K) [Y]
Soil temperature, 0.0–0.1-m depth (K) [Y]
Volumetric soil moisture content 0.0–0.1-m depth (fraction between wilting and saturation) [Y]
Water equivalent of accumulated snow depth (kg m^{-2}; i.e., mm) [Y]
2-m temperature (K) [Y]
2-m specific humidity (kg kg^{-1} dry air) [Y]
Maximum temperature (K) in last 6-h period (0000, 0600, 1200, 1800 UTC) or in last 3-h period (0300, 0900, 1500, 2100 UTC) [Y]
Minimum temperature (K) in last 6-h period (0000, 0600, 1200, 1800 UTC) or in last 3-h period (0300, 0900, 1500, 2100 UTC) [Y]
10-m u wind component (ms^{-1}) [Y]
10-m v wind component (ms^{-1}) [Y]
Total precipitation (kg m^{-2}; i.e., mm) in last 6-h period (0000, 0600, 1200, 1800 UTC) or in last 3-h period (0300, 0900, 1500, 2100 UTC) [Y]
Water runoff (kg m^{-2}; i.e., mm) [Y]
Average surface latent heat net flux (W m^{-2}) [Y]
Average sensible heat net flux (W m^{-2}) [Y]
Average ground heat net flux (W m^{-2}) [Y]
Convective available potential energy (J kg^{-1}) [Y]
Convective inhibition (J kg^{-1}) [Y]
Precipitable water (kg m^{-2}; i.e., mm) [Y]
Total-column integrated condensate (kg m^{-2}; i.e., mm) [Y]
Total cloud cover (%)
Downward shortwave radiation flux at the surface (W m^{-2}) [Y]
Downward longwave radiation flux at the surface (W m^{-2}) [Y]
Upward shortwave radiation flux at the surface (W m^{-2}) [Y]
Upward longwave radiation flux at the surface (W m^{-2}) [Y]
Upward longwave radiation flux at the top of the atmosphere (W m^{-2}) [Y]
Potential vorticity on the 320-K isentropic surface (~10^{-6} K m^2 kg^{-1} s^{-1})
U component on 2-PVU (1 PVU = 1×10^{-6} K m^2 kg^{-1} s^{-1}) isentropic surface (m s^{-1})
V component on 2-PVU isentropic surface (m s^{-1})
Temperature on 2-PVU isentropic surface (K)
Pressure on 2-PVU isentropic surface (Pa)
80-m u wind component (m s^{-1}) [Y]
80-m v wind component (m s^{-1}) [Y]
Vertical velocity at 850 hPa (Pa s^{-1})
Water runoff (kg m^{-2}; i.e., mm)
Wind mixing energy at 80 m (J) [Y]

generally in Hamill and Whitaker (2006). Skill scores were calculated in the conventional manner (Wilks 2006), ignoring the tendency to overforecast skill by not separating the data into subsets with homogeneous climatological uncertainty (Hamill and Juras 2006). Analog dates were selected on similarities of past ensemble-mean precipitation forecasts to the current ensemble-mean forecast for the current grid point and others in a ~100-km (7 × 7 grid point) box around the point of interest. Probabilities were then estimated from the ensemble of analyzed conditions for the dates with the closest match. Different numbers of analogs were used, depending on how unusual the precipitation forecast was for the day

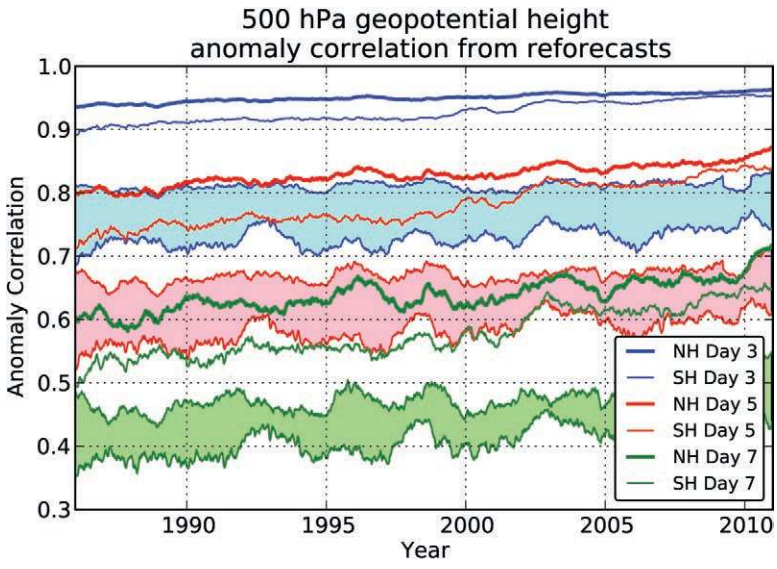

Fig. 1. Running mean (an average over the previous 365 days) of the 500-hPa geopotential height AC from the deterministic control reforecasts. The filled areas denote anomaly correlation from the first-generation GFS reforecast described in Hamill et al. (2006); the bounding lower line denotes the Southern Hemisphere AC and the bounding upper line the Northern Hemisphere AC. Blue indicates day +3 forecasts, pink indicates day +5 forecasts, and green indicates day +7 forecasts. The second-generation reforecasts are shown without filled areas; thicker lines denote Northern Hemisphere AC and thinner lines the Southern Hemisphere AC.

in question. When the event was rather common, judged relative to the forecast climatology, as many as 200 members were used. When the forecast event was in the extreme tail of the forecast distribution, as few as 30 analogs were selected. The use of fewer analogs for extreme events, especially for the short lead times, improves the forecast skill (Hamill et al. 2006, their Fig. 7). Confidence intervals were calculated with a paired block bootstrap algorithm following Hamill (1999). North American Regional Reanalysis (NARR) 24-h accumulated precipitation analysis data (Mesinger et al. 2006; Fan et al. 2006) were used both for training (cross validated by year) and verification. There are systematic errors with the NARR (Bukovsky and Karoly 2007). Still, currently we know of no other precipitation analysis that has the NARR's complete coverage of the contiguous United States over the full period of the reforecasts. We use it here, for better and worse.

The postprocessed forecasts validated from 1985 to 2010 show an improvement of slightly greater than +1 day additional lead time at the early forecast leads from the first- to the second-generation reforecast; that is, a 24–48-h version 2 forecast could be made as skillfully as the previous 0–24-h forecast from version 1. At longer leads, the improvement sometimes approaches +2 days additional lead time. All differences are statistically significant. The improvement of postprocessed forecasts from version 1 to version 2 is smaller than the improvement in the raw forecast guidance. This is to be expected; the postprocessing is correcting more systematic error in version 1 than in version 2. Postprocessed guidance from both versions is highly reliable, though forecasts from version 2 tend to issue high and low probabilities more frequently; that is, they are more "sharp" (not shown). Forecast skill probably is overestimated somewhat for the samples early on in the reforecast period (e.g., the 1980s), for the cross-validated training procedure used analogs from future forecasts that were more accurate. Experimental products based on this method are available over the contiguous United States in near–real time (at www.esrl.noaa.gov/psd/forecasts/reforecast2/analogs/index.html).

Deterministic forecasts can also be improved with the statistical postprocessing. A slightly different approach was used to generate the deterministic forecast from the analogs. First, rather than using the observed on days with similar forecasts, the difference between observed minus forecast on the days with the closest analog forecasts was used to "dress" the current forecast; this provided somewhat higher precipitation amounts when anomalously large events were forecast. The mean of this dressed set of analog forecasts was then computed. As with deterministic forecasts generated from an ensemble-mean forecast, the analog-mean forecast tends to overforecast the light precipitation and underforecast heavy precipitation. To ameliorate this, following Ebert's probability-matched mean approach (www.cawcr.gov.au/staff/eee/etrap/probmatch.html) the ensemble mean of the analogs was adjusted before it was used as a deterministic forecast. Specifically, for all the forecasts for a given month of the year, the cumulative distribution function (CDF) of these analog ensemble-mean forecasts was computed (cross validated) using the current month and the surrounding two months, as well as the CDF of the NARR dataset. The quantile associated with the current analog-mean forecast relative to the forecast climatology was noted, and the final deterministic forecast

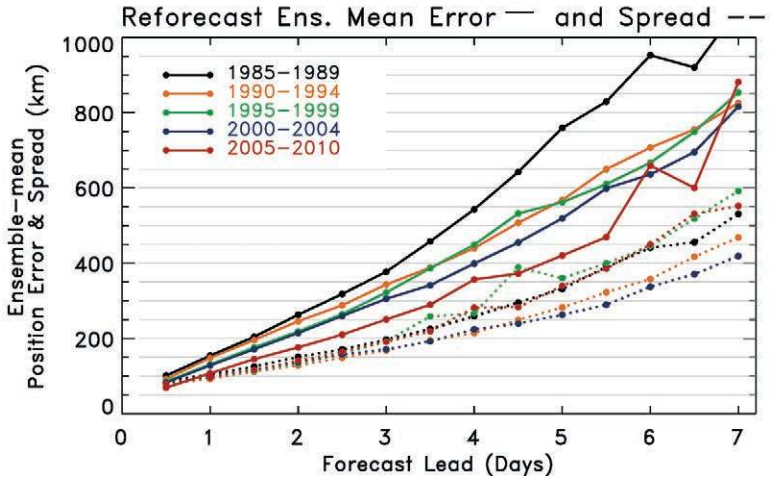

Fig. 2. Global tropical cyclone track error (solid lines) and spread (dashed) over ~5-yr periods during the reforecast. Statistics were accumulated only for 1 Jun to 30 Nov of each year and included data from all basins.

overforecast bias. Interestingly, the probability-matched mean analogs provided little improvement in skill relative to the ensemble mean or control at the longer forecast lead times. We believe that this is a consequence of applying the probability-matching process. Though this improves forecast bias, if there is little association between forecast and observed anomalies, as becomes more common at longer leads as skill degrades, then the algorithm can become overconfident of extreme events. For more on this, see Hamill and Whitaker (2006, their Figs. 2 and 7 and associated discussion).

These calibration approaches are relatively simple; they are univariate, based only on the forecast precipitation amount, and they do not factor in changes in skill of the forecasts during the training period such as may be due to increasing observational data density with time. Though not attempted here, there have been several other methods proposed in the recent past that may also be worthy of consideration, including quantile regression (Bremnes 2004), Bayesian model averaging (Sloughter et al. 2007), logistic regression (Hamill et al. 2008), and mixture models (Bentzien and Friederichs

was the precipitation amount associated with the corresponding analyzed quantile. Figure 4 shows that the analog postprocessed deterministic forecast skill also provides an improvement relative to either the GEFS control or ensemble mean, particularly at the light precipitation amounts, where apparently there was a drizzle overforecast bias. The ensemble mean from the raw ensemble shows a characteristic underforecast bias, while the control forecast has a slight

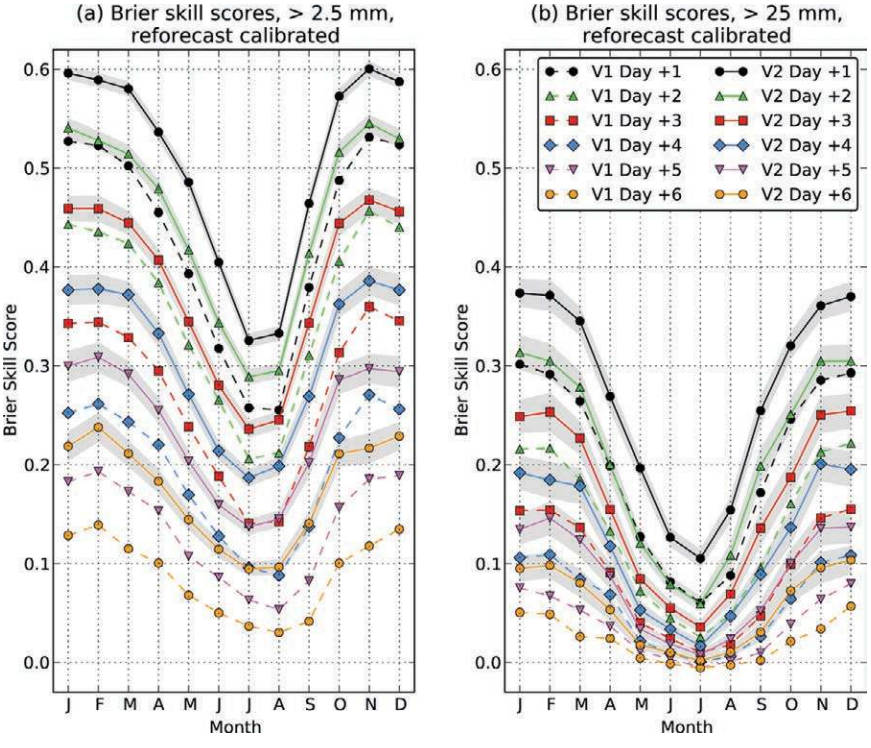

Fig. 3. Brier skill scores (BSS) of 24-h accumulated precipitation forecasts from 1985 to 2010 over the continental United States (CONUS), postprocessed using the rank analog technique. (a) BSS for the >2.5 mm 24 h–1 event. (b) BSS for the >25 mm 24 h–1 event. Scores are plotted as a function of month of the year and for different forecast lead times from 1 to 6 days. Solid lines indicate the scores for the second-generation reforecast (V2), dashed lines for the first-generation reforecast (V1). Black, green, red, blue, purple, and orange lines indicate the respective skills for days +1 to +6. Edges of the shaded gray regions provide the 5th and 95th percentiles of the confidence interval, determined via a 1000-sample paired block bootstrap following Hamill (1999).

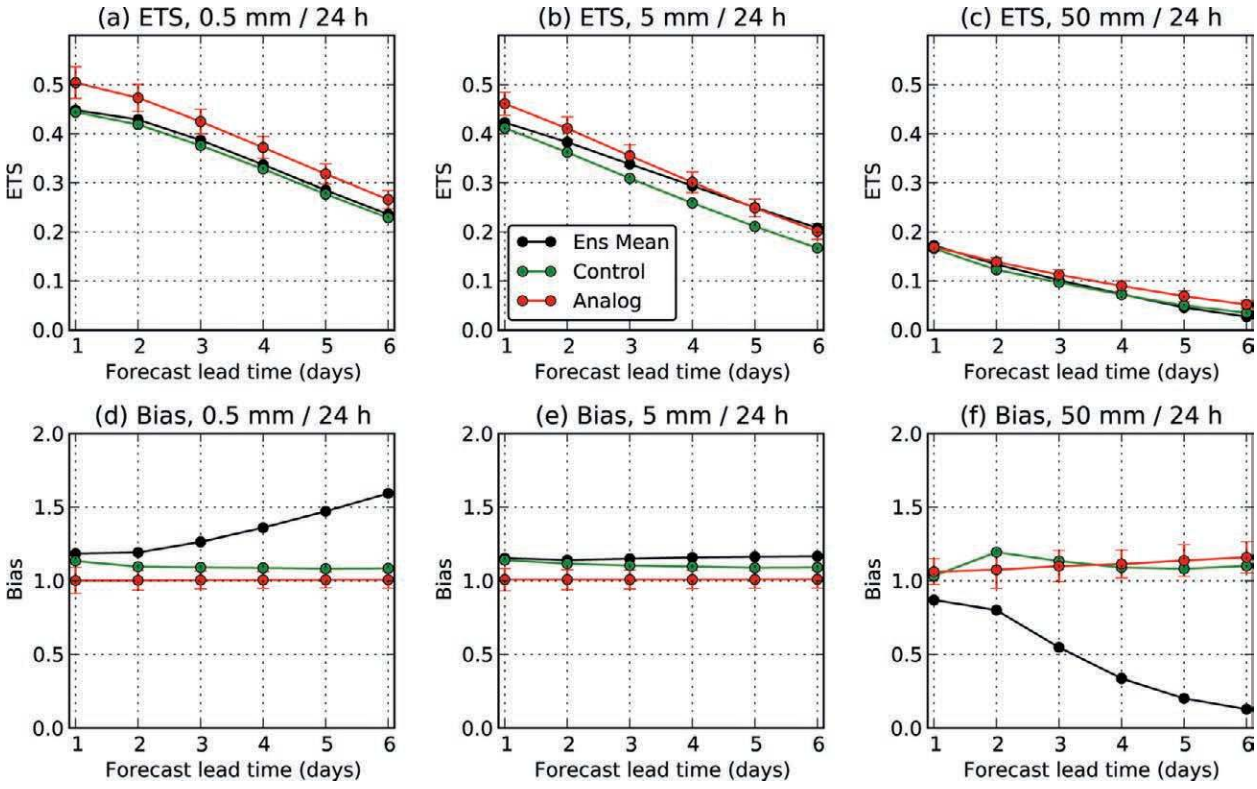

Fig. 4. Equitable threat scores (ETS) and biases (BIA) for raw ensemble-mean forecasts, control forecasts, and deterministic forecasts generated from postprocessed analog ensemble-mean forecasts. ETS for (a) the >0.5 mm 24 h^{-1} event, (b) the >5 mm 24 h^{-1} event, and (c) the >50 mm 24 h^{-1} event. (d)–(f) BIA for these respective events. The 5th and 95th percentile confidence intervals for the difference between the raw ensemble mean and the deterministic analog are plotted over the analog results. Confidence intervals were calculated with a 1000-sample block bootstrap following Hamill (1999).

2012). We hope and expect that other groups will explore methods that may extract further value from the extensive reforecast dataset, using different and new techniques and additional predictors, and test them against existing techniques. This dataset may be helpful in such comparative evaluation of different methods.

Suppose now that a long time series of observations is not available to accompany the time series of reforecasts. How can one leverage the reforecasts to provide value-added guidance? Reanalyses might be used for the calibration, but analyses may be contaminated somewhat by model forecast bias. Should the user desire guidance for a point location, the reanalysis cannot provide this, only for the gridbox averaged analyzed state. In such cases, perhaps usage of diagnostics like the extreme forecast index (EFI; Lalaurette 2003, 2013) may be of use. The EFI quantifies how unusual the current ensemble guidance is relative to the climatology of past forecast guidance. Ideally, even when the ensemble guidance is biased in some fashion, it can still provide some advanced warning of potential extreme events. For such events, today's ensemble guidance should be ranked in the extreme quantiles of the distribution defined by the past forecasts.

Figure 5 considers the problem of extended-range wind energy forecasts, specifically a +5- to +10-day forecast of 80 m above ground level wind speeds—a common height of the hubs of wind turbines. Suppose a wind farm operator in North Dakota does not have a multidecadal time series of wind observations at hub height, but he or she wishes to extract some information from a reforecast that may indicate when it would be relatively inexpensive to shut down a turbine for maintenance. Figure 5a shows the ensemble-mean forecast wind speed for a particular case day in early 2010. The winds appear relatively light on average in this location, but they might be biased. However, the availability of the reforecasts allows that wind speed forecast to be placed in context. Figure 5b shows the quantile of the ensemble-mean forecast wind speed relative to its climatology for that month—a calculation similar in spirit to the EFI. The wind speed forecasts are indeed unusually light in this location relative to their forecast climatology, which ended up being consistent with analyzed conditions (Figs. 5c,d).

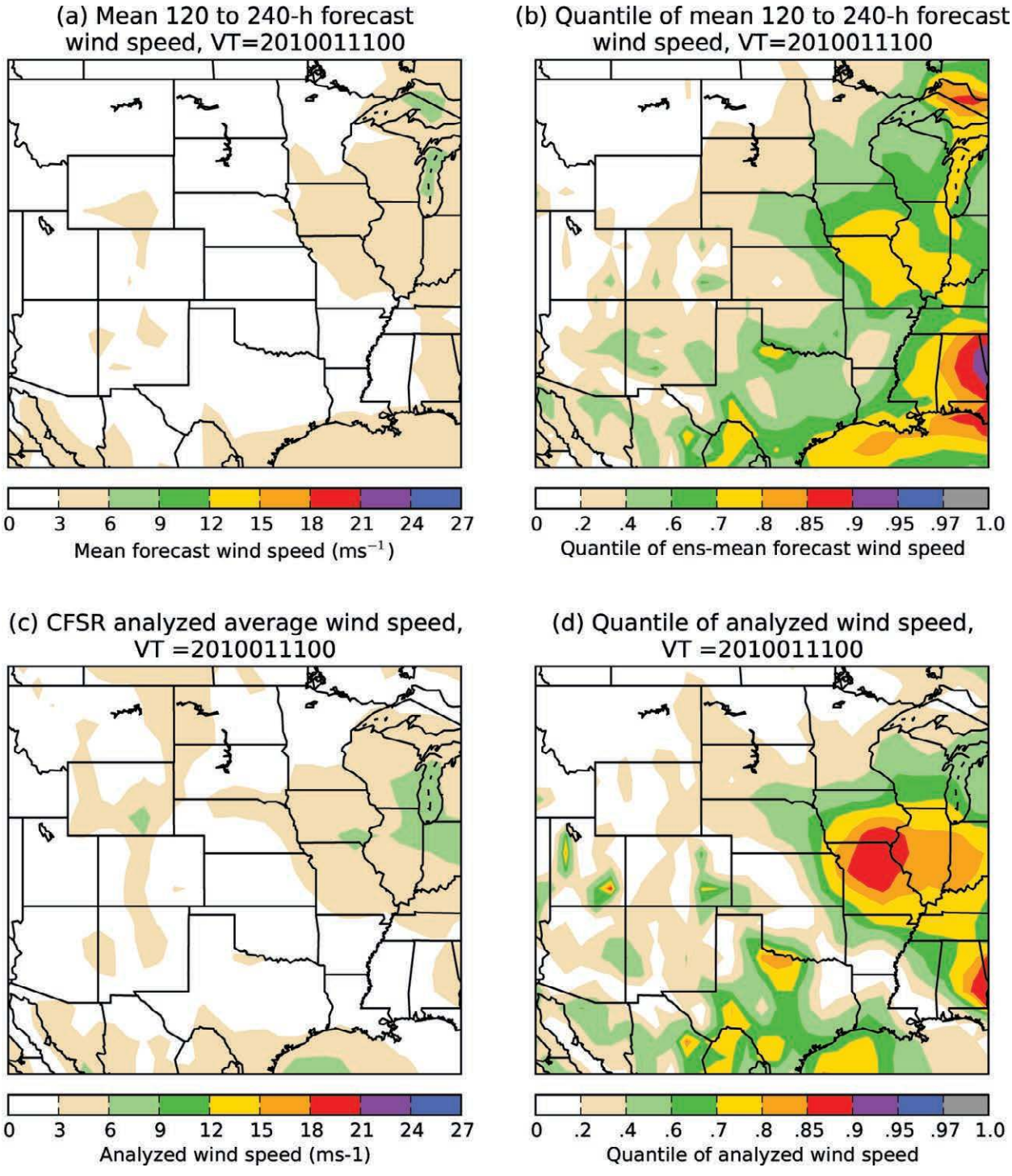

Fig. 5. (a) +5- to +10-day forecast of ensemble-mean 80-m AGL wind speeds, initialized at 0000 UTC on 1 Jan 2010 for the period 0000 UTC 6 Jan–0000 UTC 11 Jan 2010. (b) Quantile for this ensemble-mean forecast relative to the cumulative distribution of past ensemble-mean forecasts for the month of January. (c) As in (a), but for CFSR analyzed conditions, and (d) as in (b), but for CFSR analyzed.

Let us turn our attention from postprocessing to other potential applications of the reforecasts. One possible application is to use the global reforecast ensemble data as initial and lateral boundary conditions for a high-resolution regional reforecast ensemble. The ability to perform high-resolution regional reforecasts may be of interest to many, perhaps to examine the ability of a higher-resolution regional model to provide value-added guidance for high-impact weather events. As discussed previously, the full model output

for the global reforecast ensemble is available on the U.S. Department of Energy website. An illustrative example of a regional reforecast ensemble is shown in Fig. 6. Here, an 11-member ensemble 72-h forecast initialized at 0000 UTC 22 September 2005 for Tropical Cyclone (TC) Rita was generated using version 3.3 of the Advanced Hurricane Weather Research and Forecasting model (AHW), with 36 vertical levels up to 20 hPa (Skamarock et al. 2008). Details of the modification of AHW for hurricane applications are described in Davis et al. (2008). This implementation of AHW was run over a fixed 36-km domain that covers the entire North Atlantic basin, North America, and the extreme eastern North Pacific (see Galarneau and Davis 2013, their Fig. 2 and Table 1). Two-way moving nests of 12 and 4 km are located within the 36-km domain, and the movement of these nests is determined by the TC's motion during the previous 6 h. Specifics on the AHW configuration are as follows: WRF single-moment 6-class microphysics (Hong et al. 2004), modified Tiedtke convective parameterization (Zhang et al. 2011) on the 36- and 12-km domains (no parameterization on the 4-km domain), Yonsei University boundary layer scheme (Hong et al. 2006), Goddard shortwave scheme (Chou and Suarez 1994), Rapid Radiative Transfer Model (Mlawer et al. 1997), and Noah land surface model (Ek et al. 2003).

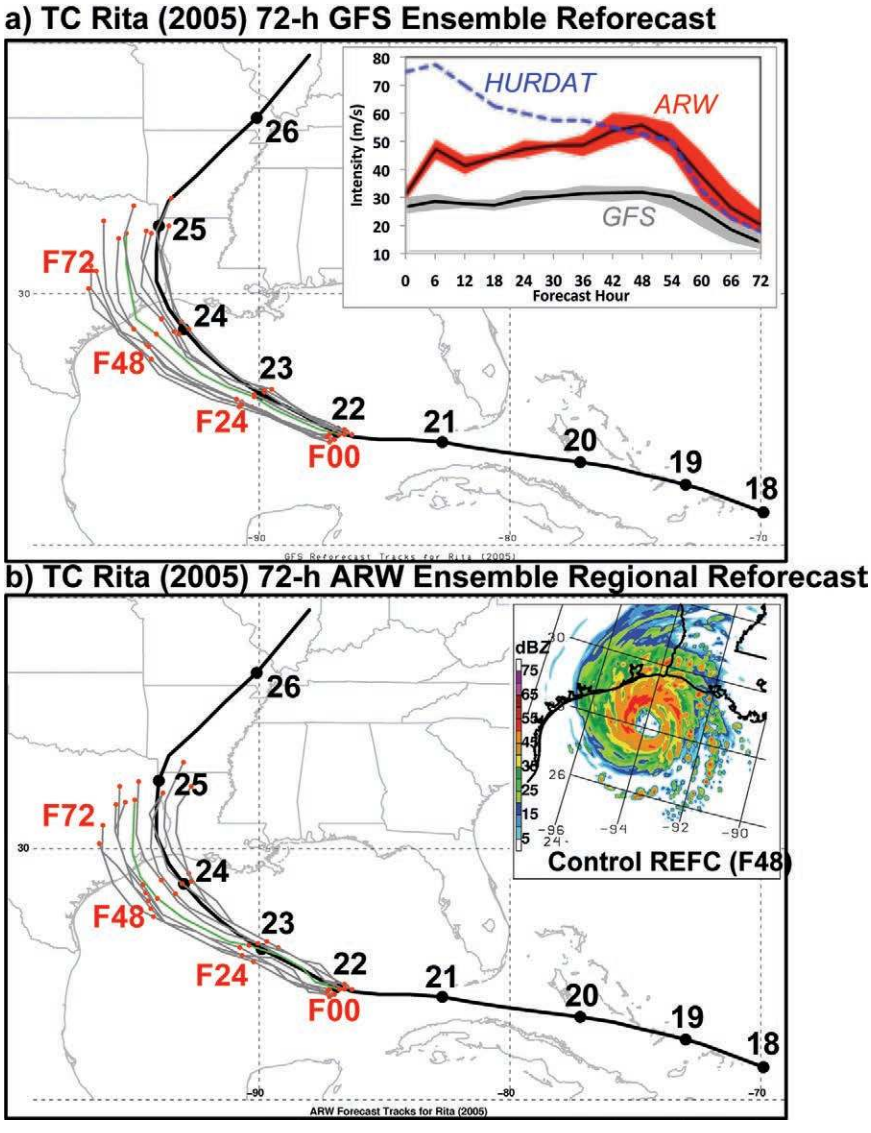

Fig. 6. A 72-h track forecast for Hurricane Rita initialized at 0000 UTC 22 Sep 2005 from the (a) global GFS ensemble reforecast and (b) regional AHW ensemble forecast. The individual ensemble member tracks are shown in gray (control run in green) with red dots marking every 24 h. The observed track is shown in black with black dots marking every day at 0000 UTC. The inset in (a) shows the intensity forecast for Rita from the global GFS ensemble (gray) and AHW (red). The observed intensity is shown by the blue dashed contour. The black line represents the ensemble mean and the shading encompasses intensity values within the 5th and 95th percentiles. The inset in (b) shows the 48-h forecast composite reflectivity (shaded according to the color bar in dBZ) from the 4-km domain of the control member of the AHW ensemble.

The global reforecast ensemble shows a range of possible model trajectories, including significant impact on Houston, Texas (Fig. 6a). The track forecast from the global reforecast ensemble was consistent with the official National Hurricane Center track forecast for Rita 3 days prior to landfall (not shown), which resulted in an evacuation order for the Houston area. The track forecast had a significant left-of-track error, as the observed storm made landfall farther northeast, near the Texas–Louisiana border. The intensity forecast was consistently underestimated in the global reforecast

ensemble (Fig. 6a inset)—a common characteristic with global data assimilation and forecast systems with grid spacing of many tens of kilometers. The AHW regional reforecast ensemble also had a left-of-track forecast error, although the ensemble track envelope expanded slightly farther northeast along the Gulf Coast (Fig. 6b). That the left-of-track error appears in the AHW reforecast ensemble in addition to the global model suggests that track errors were driven by errors in the TC steering flow. This is modulated by large-scale features such as the subtropical ridge over the U.S. Southeast and an eastward-moving midlatitude trough over the central Great Plains (not shown). The AHW reforecast ensemble inherited the initial underestimate of intensity seen in the global reforecast (Fig. 6a inset) but was able to intensify the storm to a major hurricane by 48 h, just prior to landfall (Fig. 6a,b insets).

Another potential application for reforecasts is to understand the ability of the model to predict uncommon phenomena or even the relationships between several uncommon phenomena. As an example, let us say that we wanted to understand whether atmospheric blocking statistics (Tibaldi and Molteni 1990) can be correctly forecast given a recently strong or weak MJO. To make the problem more statistically challenging, let us further suppose we are interested in the blocking forecasts related to a certain phase of the MJO, where it is most pronounced in the Indian Ocean, and at a certain time of the year, here December–February (DJF). In such a situation, a year or two of past recent forecasts will not provide enough samples.

Using the first two empirical orthogonal functions of MJO variability (Wheeler and Hendon 2004), commonly known as real-time multivariate MJO (RMM_1) and RMM_2, a strong MJO, should it exist, would be classified as being in the Indian Ocean roughly if $RMM_1 \cong 0$ and $RMM_2 \ll 0$. Accordingly, for the angle θ defined by the arctangent of RMM_1 and RMM_2, we define the Indian Ocean "strong MJO" as occurring if $-(\pi/2 + \pi/8) \le \theta \le -\pi/2 + \pi/8$, and if the amplitude $(RMM_1^2 + RMM_2^2)^{1/2}$ is in the upper quartile of the climatology of analyzed amplitudes for this phase and for DJF. Figure 7a shows the CFSR analyzed unconditional December–February 1985–2010 blocking statistics and the blocking statistics under a strong Indian Ocean MJO 6 days prior to the analysis. The lagged observed blocking frequency from the Pacific to the Atlantic Ocean is apparently strongly suppressed with strong MJOs relative to the climatology. Composites (not shown) indicate that there are generally negative 500-hPa height anomalies in the climatological

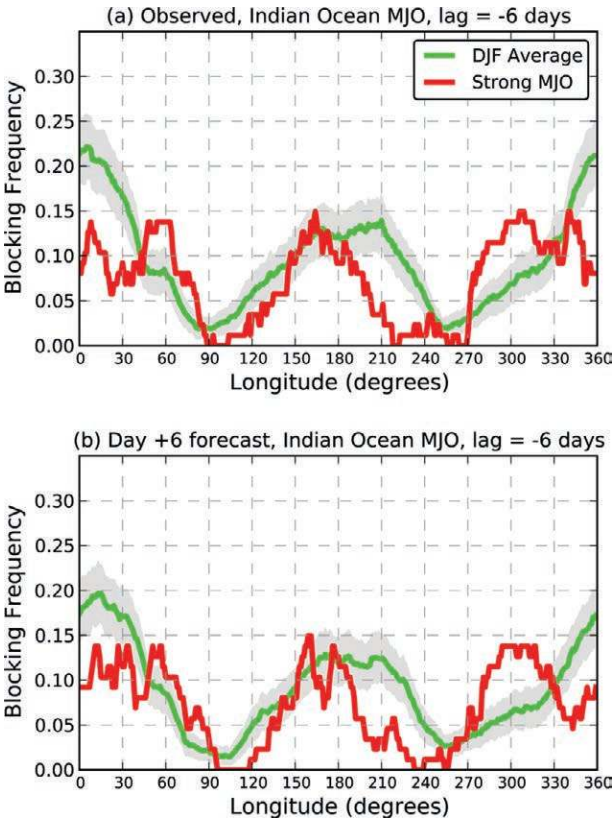

Fig. 7. (a) Observed and (b) +6-day forecast blocking frequency as a function of latitude for Dec–Jan–Feb 1985–2010 (green lines) and for the subset of cases with an Indian Ocean strong MJO as defined in the text. The MJO data were defined 6 days prior to the analysis or the forecast. Gray area denotes differences that are between the 5th and 95th percentile confidence intervals as determined from a block bootstrap algorithm.

ridges and positive anomalies in the troughs, resulting in generally more zonal flow and less blocking. Figure 7b shows the blocking frequency in the +6 day control member reforecasts (using analyzed RMM_1 and RMM_2; i.e., a –6 day lag so that analyzed data are used to define the MJO indices). There is a similar depression of the forecast blocking frequency under a strong MJO; the forecast model does well at replicating the climatology of blocking and its relationship to this phase of the MJO. This simple illustration shows how the reforecast dataset offers a unique opportunity to potentially diagnose and examine model systematic forecast characteristics related to infrequent or low-frequency phenomena.

CONCLUSIONS. For the foreseeable future, weather and climate prediction model guidance will be contaminated by at least some systematic errors. Since most end users want reliable and accurate guidance, some statistical postprocessing may be

helpful. Sometimes, such as for rare events and longer-lead forecasts, a long training dataset of "reforecasts" can be especially helpful. The large sample provides enough similar cases to statistically correct the forecasts, even with relatively uncommon events. At longer leads, the large sample can be helpful for extracting a useful forecast signal from within the bath of chaotic noise and model error (Hamill et al. 2004).

This article described one such dataset: a second-generation experimental reforecast that is approximately consistent with the 0000 UTC cycle of the NCEP Global Ensemble Forecast System as it was configured in 2012. We showed a variety of uses of this reforecast dataset, such as the statistical postprocessing of precipitation forecasts, the initialization of regional reforecasts, and the diagnosis of the forecastability of uncommon phenomena.

This dataset was generated from a large high-performance computing grant by the U.S. Department of Energy to explore the potential for improving longer-lead weather forecasts related to renewable energy; it was not created on NOAA computers. Currently, NCEP has not allocated any of its high-performance computing to the generation of reforecasts specific to weather time scales. While we intend to keep running this version of the GEFS for the foreseeable future, even after NCEP upgrades its GEFS, the regrettable truth is that soon enough the GEFS will change and the reforecast will be inconsistent with the operational version of the model. ECMWF embraced some years ago the approach of computing a more limited set reforecasts on their operational computer using whatever model version is currently operational. In this way, their reforecast dataset is continually relevant to today's model guidance. As NOAA determines the amount of high-performance computing it needs in the coming years and decades, we expect that the computers will be sized so that NOAA too can generate reforecasts (and the necessary reanalyses) regularly, save the data, and make these readily available to the weather enterprise. This current reforecast dataset will help us decide on a realistic configuration for such reforecasts.

ACKNOWLEDGMENTS. The U.S. Department of Energy provided the high-performance computing to produce this dataset under its Advanced Scientific Computing Research (ASCR) Leadership Computing Challenge (ALCC). We are grateful to the DOE and its very professional support staff for their help. The mass storage array within ESRL was partially supported by NOAA THORPEX funds distributed by NOAA's Office of Weather and Air Quality (OWAQ). We had tremendous help from the IT staff in the Physical Sciences Division at ESRL; in particular Nick Wilde, Alex McColl, Barry McInnes, Chris Kreutzer, and Eric Estes were all helpful in configuring the storage array and helping us get to voluminous reforecast data to and from it. The AHW reforecast ensemble was generated using the Bluefire supercomputer at the National Center for Atmospheric Research (NCAR). Tony Eckel and one anonymous reviewer are thanked for their careful evaluations of this manuscript.

REFERENCES

Bentzien, S., and P. Friederichs, 2012: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Wea. Forecasting,* **27,** 988–1002.

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.,* **132,** 338–347.

Bukovsky, M. S., and D. J. Karoly, 2007: A brief evaluation of precipitation from the North American Regional Reanalysis. *J. Hydrometeor.,* **8,** 837–846.

Chou, M.-D., and M. J. Suarez, 1994: An efficient thermal infrared radiation parameterization for use in general circulation models. NASA Tech. Memo. 104606, 85 pp.

Davis, C. A., and Coauthors, 2008: Prediction of landfalling hurricanes with the Advanced Hurricane WRF model. *Mon. Wea. Rev.,* **136,** 1990–2005.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.,* **108,** 8851, doi:10.1029/2002JD003296.

Fan, Y., H. M. Van den Dool, D. Lohmann, and K. Mitchell, 2006: 1948–98 U.S. hydrologic reanalysis by the Noah Land Data Assimilation System. *J. Climate,* **19,** 1214–1237.

Galarneau, T. J., Jr., and C. A. Davis, 2013: Diagnosing forecast errors in tropical cyclone motion. *Mon. Wea. Rev.,* **141,** 405–430.

Gopalakrishnan, S., and Coauthors, 2012: Use of the GFDL vortex tracker. Hurricane Weather Research and Forecasting (HWRF) model: 2012 scientific documentation. Development Testbed Center, 71–91. [Available online at www.dtcenter.org/HurrWRF/users/docs/scientific_documents/HWRFScientificDocumentation_v3.4a.pdf.]

Hagedorn, R., 2008: Using the ECMWF reforecast data set to calibrate EPS reforecasts. *ECMWF Newsletter,* No. 117, ECMWF, Reading, United Kingdom, 8–13.

——, R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multi-

model forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.,* **138,** 1814–1827.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

——, and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.,* **132,** 2905–2923.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.,* **134,** 3209–3229.

——, ——, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447.

——, ——, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.,* **136,** 2620–2632.

——, J. S. Whitaker, M. Fiorino, and S. G. Benjamin, 2011a: Global ensemble predictions of 2009's tropical cyclones initialized with an ensemble Kalman filter. *Mon. Wea. Rev.,* **139,** 668–688.

——, ——, D. T. Kleist, M. Fiorino, and S. G. Benjamin, 2011b: Predictions of 2010's tropical cyclones using the GFS and ensemble-based data assimilation methods. *Mon. Wea. Rev.,* **139,** 3243–3247.

Hong, S.-Y., J. Dudhia, and S.-H. Chen, 2004: A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Mon. Wea. Rev.,* **132,** 103–120.

——, Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.,* **134,** 2318–2341.

Hou, D., Z. Toth, Y. Zhu, and W. Yang, 2008: Impact of a stochastic perturbation scheme on global ensemble forecast. *Proc. 19th Conf. on Probability and Statistics,* New Orleans, LA, Amer. Meteor. Soc., 1.1. [Available online at https://ams.confex.com/ams/88Annual/techprogram/paper_134165.htm.]

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2009: Introduction of the GSI into the NCEP Global Data Assimilation System. *Wea. Forecasting,* **24,** 1691–1705.

Kumar, A., M. Chen, L. Zhang, W. Wang, Y. Xue, C. Wen, L. Marx, and B. Huang, 2012: An analysis of the nonstationarity in the bias of sea surface temperature forecasts for the NCEP Climate Forecast System (CFS) version 2. *Mon. Wea. Rev.,* **140,** 3003–3016.

Lalaurette, F., 2003: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quart. J. Roy. Meteor. Soc.,* **129,** 3037–3057.

——, cited 2013: Two proposals to enhance the EFI response near the tails of the climate distribution. 8 pp. [Available online at www.ecmwf.int/products/forecasts/efi_guide.pdf.]

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.,* **87,** 343–360.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmosphere: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.,* **102** (D14), 16 663–16 682.

Neumann, C. J., 1972: An alternate to the HURRAN tropical cyclone forecast system. NOAA Tech. Memo. NWS SR-62, 22 pp. [Available from the National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Rd., Springðeld, VA 22151.]

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.,* **91,** 1015–1057.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF Version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp. [Available online at www.mmm.ucar.edu/wrf/users/docs/arw_v3_bw.pdf.]

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 3209–3220.

Tibaldi, S., and F. Molteni, 1990: On the operational predictability of blocking. *Tellus,* **42A,** 343–365.

Wang, W., P. Xie, S. H. Yo, Y. Xue, A. Kumar, and X. Wu, 2011: An assessment of the surface climate in the NCEP climate forecast system reanalysis. *Climate Dyn.,* **37,** 1601–1620, doi:10.1007/s00382-010-0935-7.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus,* **60A,** 62–79.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate MJO Index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.,* **132,** 1917–1932.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 627 pp.

Zhang, C., 2005: Madden-Julian Oscillation. *Rev. Geophys.,* **43,** RG2003, doi:10.1029/2004RG000158.

——, Y. Wang, and K. Hamilton, 2011: Improved representation of boundary layer clouds over the Southeast Pacific in ARW-WRF using a modified Tiedtke cumulus parameterization scheme. *Mon. Wea. Rev.,* **139,** 3489–3513.