Short-Term Climate Extremes: Prediction Skill and Predictability

EMILY J. BECKER AND HUUG VAN DEN DOOL

Climate Prediction Center, NOAA/NWS/NCEP, College Park, Maryland

MALAQUIAS PEÑA

IMSG at Environmental Modeling Center, NOAA/NWS/NCEP, College Park, Maryland

(Manuscript received 15 March 2012, in final form 5 July 2012)

ABSTRACT

Forecasts for extremes in short-term climate (monthly means) are examined to understand the current prediction capability and potential predictability. This study focuses on 2-m surface temperature and precipitation extremes over North and South America, and sea surface temperature extremes in the Niño-3.4 and Atlantic hurricane main development regions, using the Climate Forecast System (CFS) global climate model, for the period of 1982–2010. The primary skill measures employed are the anomaly correlation (AC) and root-mean-square error (RMSE). The success rate of forecasts is also assessed using contingency tables.

The AC, a signal-to-noise skill measure, is routinely higher for extremes in short-term climate than those when all forecasts are considered. While the RMSE for extremes also rises, especially when skill is inherently low, it is found that the signal rises faster than the noise. Permutation tests confirm that this is not simply an effect of reduced sample size. Both 2-m temperature and precipitation forecasts have higher anomaly correlations in the area of South America than North America; credible skill in precipitation is very low over South America and absent over North America, even for extremes. Anomaly correlations for SST are very high in the Niño-3.4 region, especially for extremes, and moderate to high in the Atlantic hurricane main development region. Prediction skill for forecast extremes is similar to skill for observed extremes. Assessment of the potential predictability under perfect-model assumptions shows that predictability and prediction skill have very similar space–time dependence. While prediction skill is higher in CFS version 2 than in CFS version 1, the potential predictability is not.

1. Introduction

This study seeks an answer to the question ''How well can we currently predict short-term climate extremes?'' Here, ''short-term climate'' means forecasts of monthly or seasonal means at long leads—that is, not weather extremes in a 5-day forecast, and not long-term climate change. Short-term climate extremes (STCEs) have important implications for energy use, agriculture, and flood or drought preparation. In this study, we investigate our current ability to predict STCEs at lead times of one to eight months over the Americas, with the goal of understanding the strengths and weaknesses of current models, and possibly more fundamental limitations to the ability to predict. We will distinguish present-day prediction skill (what we can do now) from ''predictability'' (what we can do ultimately). To our knowledge, the notion of predictability under perfect model assumptions has not yet been applied to extremes.

Extreme climate events are a subject of increasing attention. Population density and infrastructure development in sensitive areas have led to great human and economic losses from STCEs, and positive trends in the frequency of these extremes have been detected in some areas of the globe (Easterling et al. 2000a,b, and references therein.) Short-term climate extremes in some regions have been linked to climate modes such as the El Niño–Southern Oscillation (Rasmusson and Carpenter 1982; Wolter et al. 1999), the North Atlantic oscillation (Hurrell 1995), the Pacific–North Atlantic oscillation (Wallace and Gutzler 1981), or the Madden–Julian oscillation (Madden and Julian 1972).

Corresponding author address: Emily J. Becker, National Oceanic and Atmospheric Administration, Climate Prediction Center, NCWCP W/NP5 5830 University Research Court, College Park, MD 20740-3818.
E-mail: emily.becker@noaa.gov

DOI: 10.1175/JCLI-D-12-00177.1

Most assessments of the skill of long-lead climate prediction have focused on all monthly or seasonal means (rather than extremes), using both real-time and retrospective forecasts (e.g., Chen et al. 2010; Kumar et al. 2010; Wang et al. 2010). These studies of temperature, precipitation, and SST forecasts have found that skill is not very sensitive to lead time (Barnston 1994; Livezey and Timofeyeva 2008); that is, the skill does not drop off sharply as lead times increase. Instead, skill depends much more on the target month or season.

One recent skill assessment has focused on the long-lead prediction of climate extremes. Barnston and Mason (2011) examined the International Research Institute's (IRI's) real-time prediction of seasonal temperature and precipitation extremes. Their forecasts for these extremes, defined as the 15% tails of the climatological distribution, are issued for a 3-month season at a one-half month lead. They found that forecasts for both temperature and precipitation extremes had some skill, with above-normal precipitation being overforecast and above-normal temperature being underforecast.

We have examined the prediction skill and predictability of near-surface STCEs in the region of the Americas, in the form of monthly anomalies in 2-m surface temperature, precipitation rate, and adjacent ocean sea surface temperature (SST). These fields are of wide importance to many users, and even forecasts with low levels of skill are deemed useful. We use nearly 30 years of retrospective forecasts from two versions of the National Oceanic and Atmospheric Administration (NOAA) Climate Forecast System (CFS), a ''state of the art'' coupled ocean–land–atmosphere model. As of 30 March 2011, CFS version 1 (CFSv1) has been replaced by CFS version 2 (CFSv2) as the operational model.

This paper does not focus on exactly how rare an event is, or which distribution it obeys (i.e., we do not make a fit to the generalized extreme value distribution.) The approach here is far more pragmatic, comparing forecast skill for values in a category defined as ''extreme'' (observed at some grid point during the hindcast period) to forecast skill in general.

The prediction skill for extremes is interesting not only because of its obvious societal relevance but also because it addresses a fundamental issue. To understand this issue, one needs to distinguish two scenarios. In practice, when an extreme with impacts has already occurred, the public and the media ask ''Was this predicted?'' This review of performance is valuable, and hopefully diagnostic studies of what forecasts were successful and why others failed will contribute to model improvement in the future. However, we would also like to have the ''users'' look forward—to pay attention ahead of time when an extreme is predicted. In view of this we investigate two separate scenarios: the skill of forecasts given that (at a later time) an extreme was observed, and the skill of a forecast when an extreme is predicted to occur.

One may wonder why the skill of models would be different when predicting extremes versus states near the mean; the laws of physics are no less valid in extreme situations. However, two reasons can be thought of that would make predicting extremes more difficult for models. First, extremes may stretch the validity of some of the parameterization used in models. Second, an extreme is a superposition (by Fourier transform) of all scales so as to produce, by constructive interference, an extreme in a possibly small area. As skill is notoriously scale dependent (Savijarvi 1995), it seems that it would be a tall order to get extremes perfectly right, because it would require all scales to be predicted correctly.

The data and methods used in this study are presented in section 2, including a description of the model, the notation, and the 2-m temperature, precipitation rate, and sea surface temperature verification fields. Section 3 contains some discussion of the definition of ''extreme.'' Results of forecast skill assessments are in section 4, and section 5 includes our study of the predictability of STCEs. Section 6 contains a summary and discussion.

2. Data and methods

a. Model forecasts and observations

1) MODEL REFORECASTS

Twenty-nine years of global retrospective forecasts were available for both the Climate Forecast System version 1 (Saha et al. 2006) and version 2 (S. Saha et al. 2012, unpublished manuscript). CFSv1 reforecasts cover 1981–2009, and CFSv2 covers 1982–2010. The retrospective forecasts, regridded to 1.0° longitude × 1.0° latitude, were obtained from the Environmental Modeling Center (EMC), NOAA–National Weather Service (NWS)–National Centers for Environmental Prediction (NCEP). We use monthly mean data exclusively, and initial conditions from all 12 months of the year were available.

The CFSv1 ensemble is made up of 15 members, each one a full 9-month integration starting from atmospheric initial conditions spanning three pentads in each month: the 9th to 13th, the 19th to 23rd, and the second-to-last day of the previous month through the third day of the current month (Saha et al. 2006). In this setup, the first meaningful forecast is for the next month (at lead 1 month). The atmospheric initial conditions for the CFSv1 reforecasts are from the NCEP–Department of Energy (DOE) global reanalysis 2 (Kanamitsu et al. 2002). CFSv1 uses the 2003 NCEP Global Forecast System (GFS) atmospheric model, with a resolution of T62/210 km. The

oceanic component is the Geophysical Fluid Dynamics Laboratory (GFDL) Modular Ocean Model, version 3 (MOM3), which extends from 74°S to 64°N, with a zonal resolution of 1°, and a meridional resolution of 0.33° between 10°S and 10°N, increasing gradually to 1° beyond 30°S and 30°N. Ocean initial conditions are from the Global Ocean Data Assimilation (GODAS). The land surface model is the two-level Oregon State University model. For more details on the CFSv1, see Saha et al. (2006), and references therein.

CFSv2 contains 24 members (28 for the November initial conditions), each a full 9-month integration (S. Saha et al. 2012, unpublished manuscript). The reforecasts are initiated every fifth day, for all four cycles of each day (0000, 0600, 1200, 1800 UTC). The initial conditions for atmosphere, ocean, and land were generated by a coupled reanalysis named CFSR (Saha et al. 2010). The GFS resolution was increased to T126–100 km, and the ocean model was upgraded to MOM4, which is fully global. The horizontal resolution of the ocean model between 10°S and 10°N is 0.25°, and 0.5° elsewhere. The land surface model used by CFSv2 is the four-level Noah land model (Ek et al. 2003; Mitchell et al. 2004). For more details on the CFSv2, see Saha et al. (2010) and S. Saha et al. (2012, unpublished manuscript).

2) OBSERVATIONS

The 9-month lead forecasts initialized in 2010 stretch into 2011, and so scores for 2010 include verifying data through much of 2011. The verification field for 2-m temperature (T2m) is the station observation-based Global Historical Climatology Network version 2 and the Climate Anomaly Monitoring System (GHCN+CAMS; Fan and van den Dool 2008). This global land monthly mean surface air temperature dataset is a combination of two large individual datasets of station observations, and tests have found that most common temporal–spatial features are captured (Fan and van den Dool 2008). GHCN+CAMS has a native resolution of 0.5° latitude × 0.5° longitude and was regridded to 1.0° × 1.0° for this study.

The precipitation rate was examined using the Climate Prediction Center (CPC) global daily Unified Rain-Gauge Database (URD) gauge analysis for verification. This global land-only dataset uses quality-controlled input from over 30 000 stations in the Global Telecommunication System (GTS) and many other national and international collections (P. Xie et al. 2010, unpublished manuscript). The URD is also available on a 0.5° latitude/longitude grid and was regridded to 1.0° × 1.0° for this study. Monthly means were prepared from the daily data.

The sea surface temperature verification data (often called OI-2) is that of Reynolds et al. (2002) and uses both satellite data and in situ records from ships and buoys. It is an optimum interpolation analysis, produced at NOAA. The native resolution of the Reynolds et al. (2002) SST is 1° latitude × 1° longitude.

b. Notation

The monthly mean data we use throughout can be represented by $X(s, j, m, \tau)$, where s is a spatial index, j stands for the year (1982–2010), m is the target month (1–12), and τ is the forecast lead in months (0–8). The quantity X is either the forecast F or the observation O, and both F and O may refer to temperature, precipitation rate, or sea surface temperature. For the observations, the notation $O(s, j, m)$ suffices.

Anomalies, denoted by the prime ($'$), are formed by

$$O'(s,j,m) = O(s,j,m) - C_O(s,m), \tag{1a}$$

where $C_O(s, m)$ is the observed climatology calculated over many years (often 1982–2010, but not necessarily). Likewise, we form forecast anomalies by

$$F'(s,j,m,\tau) = F(s,j,m,\tau) - C_O(s,m), \tag{1b}$$

where we note, with emphasis, that the same $C_O(s, m)$ is subtracted in (1a) and (1b). Obviously, $F(s, j, m, \tau)$ could have a systematic error, and an attempt to correct this error can be interpreted as an adjustment to Eq. (1b); this will be discussed below in section 2c.

ENSEMBLE FORECAST NOTATION

Forecasts actually come as an ensemble of forecasts, so adding one more argument is necessary to be complete for the forecast notation: $F(s, j, m, n, \tau)$, where n is the ensemble number, $n = 1 \ldots N$, where $N = 24$ (for CFSv2) in most months. All the expressions given above would apply to either the verification of a single forecast, or the ensemble mean, defined as $F_{\text{ens}}(s, j, m, \tau) = \sum_n F(s, j, m, n, \tau)/N$. The availability of ensemble members also gives us room to experiment with the definition of ''extreme'' from a model standpoint. For instance, we could label cases extreme when abs(F'_{ens}) is above a certain threshold, or perhaps when k ensemble members are larger than some threshold. This will be explored further in section 3.

c. Systematic error correction

Long lead forecasts based on dynamical models are often plagued by systematic errors (SEs), and given low to moderate intrinsic skill we need to do a SE correction (SEC) to bring out the skillful part of the forecast. Identifying and correcting systematic bias in the models often leads to an improvement in forecast skill (Smith and Livezey 1999). The generic definition of the SE is the difference in the expected value of the forecasts and

verifying observations; this is estimated as the mean over as many cases (years) as possible. In view of Eq. (1b), it would be good to bias correct F; that is,

$$F'(s, j, m, \tau) = F(s, j, m, \tau) - \{F(s, j, m, \tau)\} - \{O(s, j, m)\} + C_O(s, m), \quad (2)$$

where $\{X\}$ is the mean over many years and the $\{F - O\}$ term is our estimate of the SE. It is easy to see that (2) or (1b) reduces to $F'(s,j,m,\tau) = F(s,j,m,\tau) - \{F(s,j,m,\tau)\}$ if the years involved in calculating $C_O(s, m)$ and $\{O(s, j, m)\}$ are identically the same, for example when 1982–2010 is used for both. However, this shortcut, while commonly done, is not a good practice for many reasons. For instance, when we do a cross validation (CV) leaving three years out, the SEC [i.e., $\{F(s, j, m, \tau)\} - \{O(s, j, m)\}$] is evaluated over all years except three (Barnston and van den Dool 1993). Therefore, the external climatology, $C_O(s, m)$, needs to be shown explicitly in Eq. (2). With regard to CV, we have decided that three years need to be left out at a minimum. The procedure is referred to as CV3RE, meaning three years are left out, the test year and two more years chosen at random (the R), and the E refers to an ''external'' climatology. Ideally $C_O(s, m)$ is taken from observations entirely outside the dataset under analysis (here 1982–2010). But we do not often have that luxury, and even if we had observational data for, say, 1850–1980, we would have to worry about climate change becoming a factor. The next best thing is to keep $C_O(s, m)$ fixed and not change it in response to certain years being left out under a CV. More background and justification for CV3RE can be found online (at http://www.nws.noaa.gov/ost/climate/STIP/FY09CTBSeminars/vandendool_051109.htm).

d. Verification measures

There are many methods of assessing forecast skill. In this study, we mainly rely on the anomaly correlation (AC) and root-mean-square error (RMSE) to assess the skill of forecasts in general, to assess the skill of STCE only (a subset of all forecasts), and to calculate predictability. We also examine the forecast ''hits and misses'' using contingency tables, detailed in section 4.

The AC is a measure of the association between the anomalies of (usually) gridpoint forecast and observed values (Wilks 1995; van den Dool 2007) and is given by

$$\mathrm{AC}(m, \tau) = \frac{\sum_s \sum_j F'(s,j,m,\tau) O'(s,j,m)}{\left\{ \sum_s \sum_j [F'(s,j,m,\tau)]^2 \sum_s \sum_j [O'(s,j,m)]^2 \right\}^{1/2}}, \quad (3a)$$

where the double summation is over years (j = 1982–2010) and space (e.g., all land grid points over North America if 2-m temperature or precipitation is considered). A weight, not shown, may be carried to account for the area represented by each grid point. The expression in the numerator is a covariance between F and O, while the denominator carries two standard deviations. The correlation coefficient is denoted by AC, a number between −1 and +1, where +1 refers to a perfect forecast and 0 to random forecasts. Negative values for AC may occur when there is little or no skill, and a sample small enough for nonnegligible sampling variability. The summation over space can be suppressed to yield a spatial distribution of the AC [i.e., $\mathrm{AC}(s, m, \tau)$], given by

$$\mathrm{AC}(s, m, \tau) = \frac{\sum_j F'(s, j, m, \tau) O'(s, j, m)}{\left\{ \sum_j [F'(s, j, m, \tau)]^2 \sum_j [O'(s, j, m)]^2 \right\}^{1/2}}. \quad (3b)$$

A second verification measure is the root-mean-square error, defined as

$$\mathrm{RMSE}(m, \tau) = \left\{ \sum_s \sum_j \frac{w_s [F(s, j, m, \tau) - O(s, j, m)]^2}{W} \right\}^{1/2}, \quad (3c)$$

where $W = \sum_s \sum_j w_s$.

Any findings of skill for predicting extremes will depend strongly on how skill is defined, and here the prediction of extremes may benefit from what we usually call skill. For instance, the anomaly correlation is a prime example of a signal-to-noise measure, while RMSE is a measure of the noise only. Conclusions based on root-mean-square error may thus be different from those based on the anomaly correlation, which thrives on high signals, especially those of extremes. Hence, we examine both the AC and RMSE measures.

e. Predictability

While Eqs. (3a)–(3c) evaluate prediction skill of a model against observations (normal procedure) there is also the notion of ''predictability.'' There are various methods of defining predictability; one common definition is evaluated as one model forecast versus another (Lorenz 1982; NRC 2010). This requires an ensemble, and we will evaluate predictability as the AC achieved by an ensemble mean over $N - 1$ members, and using the one member left out as the substitute observation. In this context one makes the so-called perfect model

assumption (i.e., we know for sure that the forecast and proxy-observation are taken from the same world and there are no systematic errors to be corrected for). The expression is as follows:

$$\mathrm{AC}_p(m,\tau) = \frac{\sum_s \sum_j F'_{\mathrm{ens}}(s,j,m,\tau)F'(s,j,m,n,\tau)}{\left\{\sum_s \sum_j [F'_{\mathrm{ens}}(s,j,m,\tau)]^2 \sum_s \sum_j [F'(s,j,m,n,\tau)]^2\right\}^{1/2}}, \quad (4)$$

with the proviso that F_{ens} is calculated from $N - 1$ members, and $F'(s, j, m, n, \tau)$ is the one member (the nth member) left out. (One can repeat this calculation N times, depending on which member is left out, expecting the same answer in each case, except for sampling error.) The anomalies are formed relative to our best estimate of climatology in the model world [i.e., $\sum_j \sum_n F(s, j, m, n, \tau)/(JN)$, where J is the number of years]. In the above we assume all members to be ''equal.'' To control for small variations in predictability, the calculations in this study are performed for each of the N member, and the results for all the ensemble members were averaged together.

3. The definition of "extreme"

There are many ways of defining an extreme, including values of a variable that fall above or below a local threshold, the tails of various frequency distributions, or other characteristics, and this definition may affect the outcome of a study. For example, Sheridan and Dolney (2003) examined heat waves in Ohio using four different criteria for identifying a heat event, and found that the increase in mortality levels varied somewhat depending on the criteria. The variety of definitions has complicated studies of trends in extremes (Nicholls 1995; Easterling et al. 2000b). Also, statistical definitions of ''extreme'' carry no guarantee that a local extreme has a significant effect on residents, agriculture, or resources. However, especially for a study of a large geographical area, we must use some criteria that can be universally applied. For the purposes of this study, we have defined a climate extreme as a departure from the monthly mean above/below a specified multiple of the local standard deviation of the variable. Specifically, we use standardized anomalies, where the anomalies generated using Eq. (1) are divided by the gridpoint standard deviation, and define ''extreme'' as a standardized anomaly usually as above/below ±1.645. This is approximately equivalent to the 5th and 95th percentile threshold, and, in a record of 29 years, results in around 2–4 extreme cases per grid point. In some figures we use a continuous sliding scale so the reader can decide what is extreme. We will examine the results of skill analysis for some variations on this criterion in section 6. While it is unlikely that every identified extreme for each grid point is an event of great impact, we do trust that all large impact events are included. Our method ensures a dataset large enough for a meaningful statistical analysis.

Observed extreme versus predicted extreme

Not all forecasts for extremes come true, and not all observed extremes were forecast. This perspective leads one to consider whether the skill for the situation of verifying a forecast against an observation taken at a later time should be any different from the skill of verifying an observation against a forecast made earlier. The expressions we use for assessing error, the RMSE and AC, are ''symmetric'' in terms of the forecast F and the verifying observation O. For instance, see the expression we use for RMSE, Eq. (3c) above. Fundamentally, there is no difference between these two situations; this is the symmetry. As long as *all cases* are included, the RMSE in (3c) is the same for both settings. (This also holds for the anomaly correlation.) When we verify only over a few cases, this symmetry could possibly be broken, since, for the same time and grid point, F could be an extreme and O not (and vice versa). Consider two different scenarios for limiting the input to the AC and RMSE calculations: on the one hand, those cases conditioned on an *observed* extreme having occurred, and, on the other, those cases conditioned on the *forecast* being an extreme. For such subsets Eqs. (3a)–(3c) may not yield the same answer. An additional asymmetry arises when the criterion for ''extreme'' is based on an ensemble of F realizations (not a single forecast), while observations occur only once.

4. Prediction of monthly temperature, precipitation, and sea surface temperature

The performance of the model is assessed in a few different ways. For each variable, three scenarios were assessed: 1) the prediction skill of the models over the entire 29-yr record; 2) when an extreme event occurred in the observed record (i.e., an extreme occurred, was it

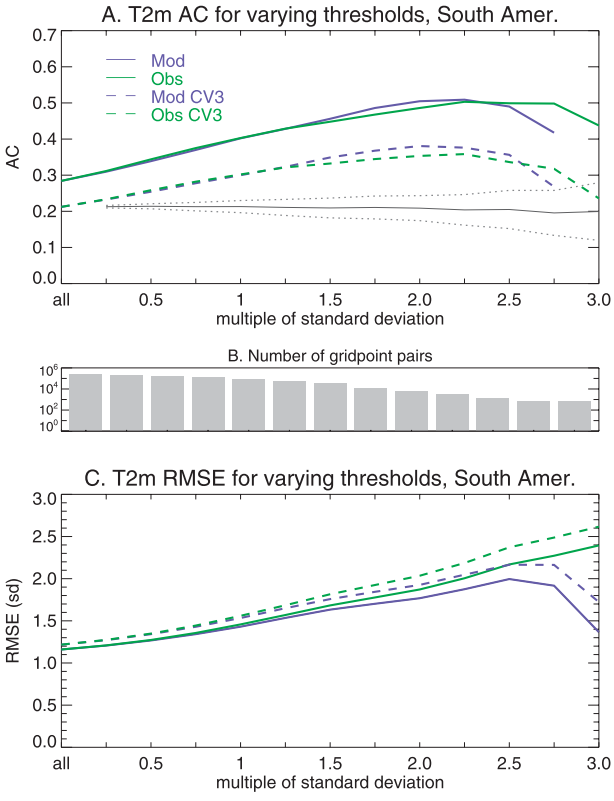

FIG. 1. (a) Two-meter temperature anomaly correlation (AC) and (c) root-mean-square error (RMSE, in units of standard deviation) for CFSv2 predictions over South America. Each value is the aggregate of the leads 1–3 months AC or RMSE over all initial conditions in all years. Horizontal axis indicates the multiple of the standard deviation that was used to filter the data sample. (b) On a logarithmic scale, the number of gridpoint pairs involved in the calculation, decreasing to the right by orders of magnitude. Purple and green lines show the AC–RMSE values for extremes identified in the model ensemble mean and the observations, respectively. Dashed lines show the values for cross validation when three years are excluded (CV3RE, see text for discussion.) The gray lines in (a) depict the mean (solid) and ±2 standard deviations (dashed) from the mean of the AC calculated for 1000 permutations of a randomly selected, cross-validated sample with the number of gridpoint pairs shown in (b).

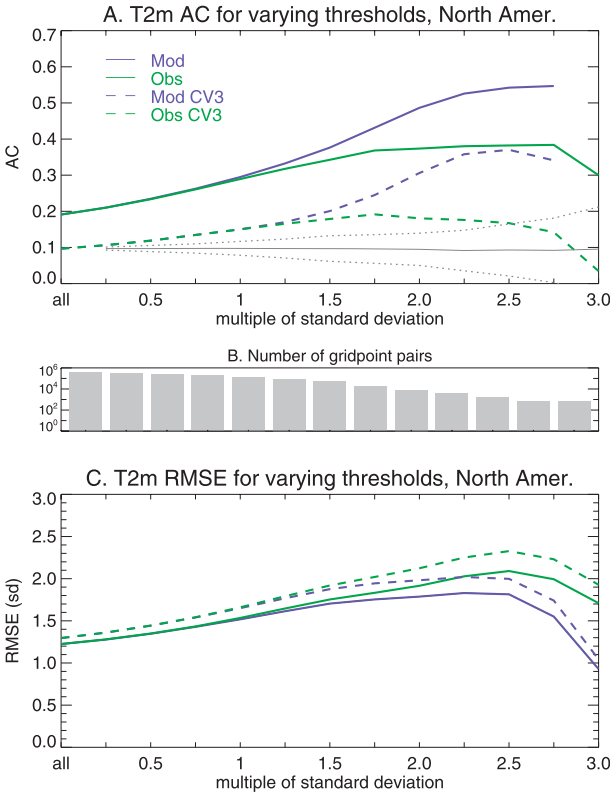

FIG. 2. As in Fig. 1, but for North America.

predicted ?); and 3) verification of a predicted extreme (i.e., the model predicted an extreme, did one occur ?). For brevity, we will refer to these respectively as ''all cases,'' ''observed extremes,'' and ''predicted extremes'' in the following discussion.

We have defined a climate extreme to be, for each grid point, an anomaly in the monthly mean above–below a threshold (see section 3 above). To study the effects of the choice of threshold, the AC and RMSE were examined for several reduced samples. For example, if a climate extreme is designated to be an anomaly above/below ±1.645 standard deviation, all anomalies that are less than the threshold are excluded from the calculation. Hence, the sample size is reduced with each larger threshold. Since we are assessing the model performance for both observed extremes and predicted extremes, the skill is tested for two separate scenarios: 1) the time series of observation–forecast gridpoint pairs is restricted to those where the *observed* value passes the criteria for an extreme, and 2) it is restricted to those where the *model-predicted* value passes the criteria.

In the following figures, area-aggregated anomaly correlations are shown. The steps to arrive at this value for each initial month are as follow: first, the anomalies for the observations are generated as in Eq. (1a), and standardized by the local gridpoint standard deviation. Bias-corrected anomalies for the model ensemble mean (EM) are generated using Eq. (2), both with and without cross validation, and standardized by the EM gridpoint standard deviation. Then the AC is calculated as in Eq. (3a) for the set of observed–forecast gridpoint pairs (either the entire set or one of the extremes scenarios), including the double summation over the 29 years and the designated spatial domain; the final step is the multiplication and division of the three summation terms. This leaves us with nine ACs, one for each lead. The RMSE is arrived at following a similar method. This

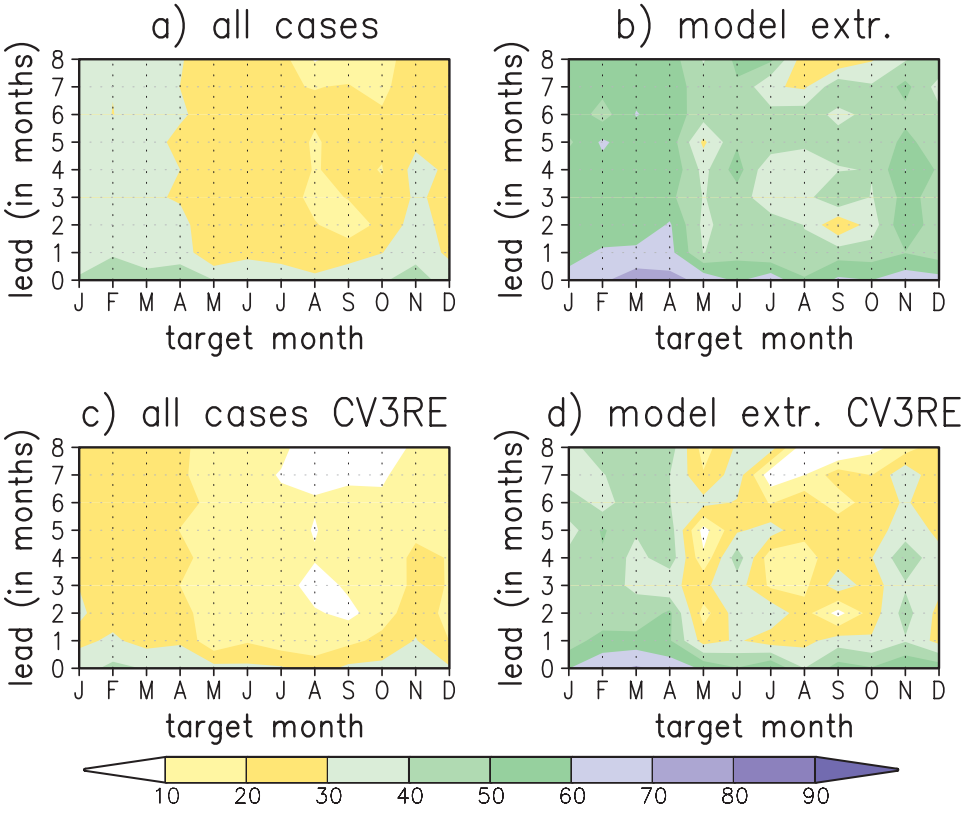

FIG. 3. CFSv2 2-m temperature anomaly correlations over South America, expressed as a function of the target month (horizontal) and lead (vertical) for (a) all cases. (b) When the model predicts an extreme, defined as ±1.645 local standard deviation. (c) All cases with CV3RE applied. (d) Predicted extremes with CV3RE applied.

process is repeated for each of the 12 initial months. The uncertainty (sampling error) in a correlation (for small correlation) is $1/\sqrt{N_{\text{eff}} - 2}$, where N_{eff} is the effective number of cases. For example, a 0.4 correlation is locally marginally significant for a single point time series over 29 years. Aggregating over large spatial domains leads to larger N_{eff}, and increased statistical significance for the same 0.4 value (Saha et al. 2006).

The anomaly correlation scores for leads 1–3 combined allow for assessment of all three scenarios (all cases, observed extremes, and predicted extremes). Leads 1–3 are chosen for the average because it is assumed that the model will have the highest performance during the early leads. Lead 0 is omitted because the CFS lead 0 is initialized with the atmospheric conditions, leading to high scores due to short-term weather forecasting, which is not the topic of this paper. Leads 1–3 forecasts from all initial months over all years are included in the average. ''South America'' is the area average for all land in the South American continent area-averaged south of 15°N. ''North America'' is area-averaged north of 15°N; Greenland is not included.

To understand how the definition of ''extreme'' affects the AC, we assess the AC when the sample is limited by many different thresholds (i.e., an increasing multiple of the local standard deviation). Figures 1 and 2 show the area-averaged AC and RMSE for leads 1–3, for 2-m temperature over South and North America, respectively. Purple lines show the results for predicted extremes, and green lines the results when the sample is filtered to observed extremes. We used the following multiples of the standard deviation in this analysis: 0 (all cases), 0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, and 3.0. The analysis was performed both without cross validation (solid lines) and with (dashed lines). Extremes of either sign are combined in this figure.

One general conclusion that will be drawn from these figures is that ACs for the extremes scenarios, both for observed and predicted extremes, are uniformly higher than for the all-cases scenario. One may wonder if this increase in AC when the sample is limited to extremes is possibly a response to the reduction in sample size. To assess this, for each threshold we examined the AC for 1000 randomly selected, cross-validated subsamples that

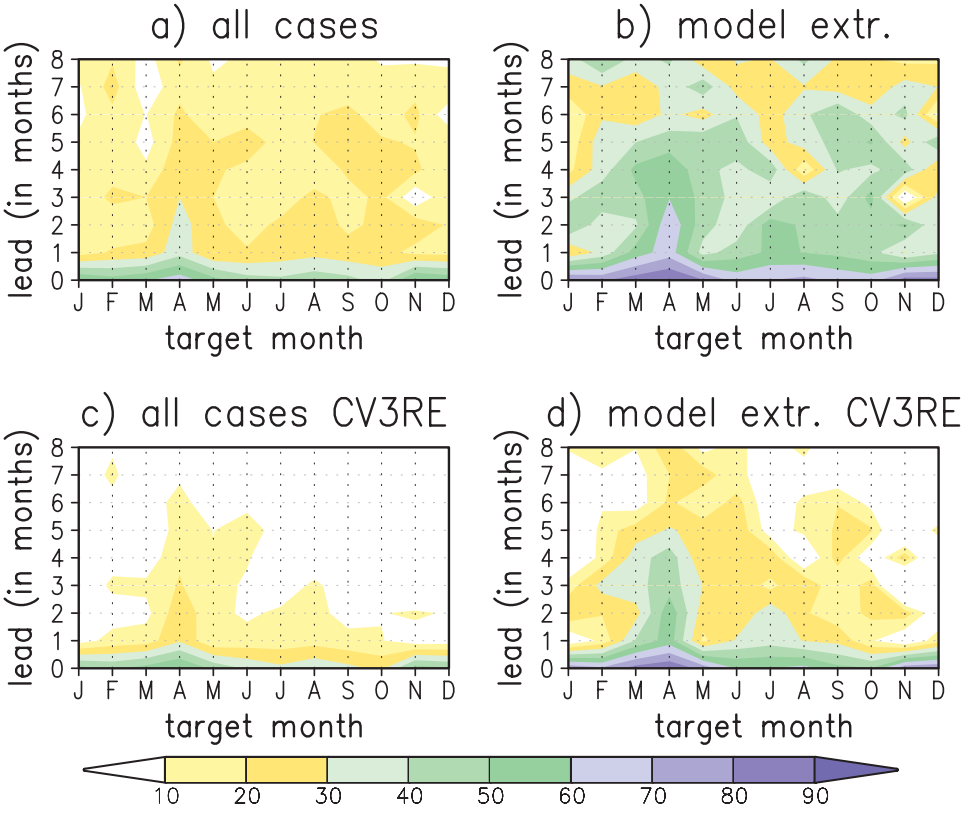

FIG. 4. As in Fig. 3, but for North America.

were the same size as the extreme subsample. The upper and lower bounds of the envelope defined by the mean ±2 standard deviations are shown as gray dotted lines; the mean of the 1000 permutations is shown by the solid gray line. Therefore, there is no question that, by the AC measure, extremes are better predicted than all cases over both Americas.

a. Prediction skill: 2-m temperature

In South America (Fig. 1a), 2-m temperature ACs for both observed extremes (green) and predicted extremes (purple) in the non-cross-validated set increase from around 0.29 to approximately 0.5 for large extremes. ACs for the sample after CV3RE has been applied (dashed lines) increase from slightly higher than 0.2 to around 0.35. Skill appears to fall off for standardized anomalies greater than 2.5, when the sample becomes very small, with <1% of data left. The results of the permutation tests on randomly selected, cross-validated samples (gray solid and dashed lines) show that the AC does not grow simply with "just any" reduced sample size, but in fact remains close to the value for all cases. The RMSE in units of standard deviation (Fig. 1c) increases slowly with the reduced sample size (shown in Fig. 1b on a log scale), and cross validation results in only slightly higher RMSE. Significantly, the slope of the RMSE lines in Fig. 1c is less than 45°; the importance of this is discussed below.

Over North America (Fig. 2a), skill when the model predicts an extreme is higher than skill when an extreme occurred in the observed record, when "extreme" is defined as a standardized anomaly greater than 1 (1.25 when cross validation is applied). The gap in ACs is larger in the CV3RE set. When the forecast is above a threshold of 2.5 standard deviations (sd), the sample is too small for significant results. As with South America, the RMSE in North America (Fig. 2c) increases slowly with the threshold, from RMSE = 1.25sd at 0 threshold to RMSE = (1.75 to 2.0)sd at the largest thresholds, just before the reduced sample size becomes an issue. When extremes are defined as a standardized anomaly above about 1.5, the RMSE in the observed extremes sample increases faster than in the predicted extremes sample. ACs, in general, are lower in North America than over South America, and the RMSE is higher. Still, the behaviors in Figs. 1 and 2 have much in common.

It may seem peculiar that both AC and RMSE go up with increasing threshold. Both are related to skill, but AC goes up *despite* the RMSE going up. The explanation for this is as follows: one can approximate AC (as

long as it is not too large) as the signal-to-noise ratio (van den Dool and Toth 1991; Compo and Sardeshmukh 2004), where RMSE is noise and the signal is at least the threshold. As one can see from Figs. 1c and 2c, RMSE increases more slowly than the threshold (the slope of the line is less than unity), and so the AC increases despite the RMSE increase.

Figures 3 and 4 depict the anomaly correlation as a function of the lead (vertical axis) and target month (horizontal axis); forecast skill is often more strongly related to the target month than to the initial month (Barnston 1994). This type of display allows us to understand what months the models are best at forecasting, and to study the forecast skill at longer leads. Here, extremes are defined as a standardized anomaly greater–less than ±1.645; this is approximately the 5th percentile anomalies (i.e., 10% of all cases). The anomaly correlations when the samples are restricted to model-predicted extremes are similar enough to those when only observed extremes are considered; hence, for simplicity, only the predicted extremes cases are shown. In South America (Fig. 3), the highest skill in 2-m temperature is during forecasts of January–April, for both the all-cases (Figs. 3a,c) scenario and for the predicted extremes (Figs. 3b,d). However, ACs over 0.2 are present for most forecast months for CFSv2, even out to the 8-month lead. September forecasts have the lowest skill. ACs for the cross-validated set (Figs. 3c,d) are generally about 0.1 point lower than for the non-CV3RE set. When judged by the AC, extremes are much better predicted than anomalies in general.

Skill over North America for all cases (Figs. 4a,c) is lower at longer leads, especially when CV3RE is applied, when low anomaly correlations are found for most target months at leads of greater than about 2 months. This is because, with a sample of 29 years, a correlation of 0.2 may not be distinguishable from 0. However, April and May do have somewhat higher scores. When predicted extremes are considered (Figs. 4b,d), the results become noisier, but ACs above 0.3 are found for leads of 5–6 months for most target months in the non-CV3RE set (Fig. 4b). These are lower in when cross validation is applied (Fig. 4d), with low scores after around 3-month leads, except for target months in the boreal spring, when scores are higher.

The outcome of dichotomous forecasts, such as a forecast for an extreme event (the event either occurs or does not occur), can be summed up using a contingency table (Wilks 1995). Four outcomes are possible: an event is forecast and occurs (''hit''), an event is forecast and does not occur (''false alarm''), an event occurs that was not forecast (''miss''), and a ''correct negative.'' For this analysis extremes are defined in both the observations and forecasts using a threshold of ±1.645 times the standard deviation. From here, the analysis is straightforward. Hits, misses, false alarms, and correct negatives are counted, for two dichotomous forecasts treated separately: positive extreme and negative extreme.

TABLE 1. Hit rate (HR), false alarm rate (FAR), and bias ratio for 2-m temperature, precipitation rate, and SST. Results are shown for forecasts for positive and negative extremes from CFSv2.

	+ extreme			− extreme		
Tmp2m	HR	FAR	BIAS	HR	FAR	BIAS
South America	0.18	0.05	1.06	0.13	0.04	0.9
North America	0.21	0.05	1.14	0.18	0.04	0.87
	+ extreme			− extreme		
P rate	HR	FAR	BIAS	HR	FAR	BIAS
South America	0.1	0.06	0.88	0.1	0.04	1.69
North America	0.08	0.05	0.79	0.06	0.05	2.8
	+ extreme			− extreme		
SST	HR	FAR	BIAS	HR	FAR	BIAS
Niño-3.4	0.44	0.02	0.84	0.44	0.03	1.21
MDR	0.42	0.03	1.01	0.18	0.03	1.03

The number of hits, divided by the total number of observed extremes (hits + misses), is the hit rate (HR); an HR of 1 would mean every forecast verified. Conversely, the number of false alarms, divided by the total number of observed nonextremes (false alarms + correct negatives) is the false alarm rate (FAR). If we are using the threshold of 1.645 standard deviations, meaning approximately 5% of forecasts are positive extremes, we can expect, in the absence of any skill, a baseline hit rate of about 0.05 for positive extremes. For the same reasons, the FAR would also be about 0.05. Bias in the distributions could affect these estimates somewhat.

A third measure, the bias ratio (B), tells if extremes are forecast more often than they occur ($B > 1$) or less ($B < 1$). These three measures together represent the information contained in a contingency table (Stephenson 2000). Table 1a shows the HR, FAR, and B values for the 2-m temperature in CFSv2 over the Americas, averaged over leads 1–3 and all initial months. The results are presented here without cross validation; as in the anomaly correlations, cross validation reduces the HRs for 2-m temperature and precipitation and has little effect on SST HRs. The 2-m temperature hit rates are higher for warm extremes than for cold over both South and North America. The bias ratio indicates that warm extremes are forecast slightly more often than they occur, while cold extremes are slightly underforecast.

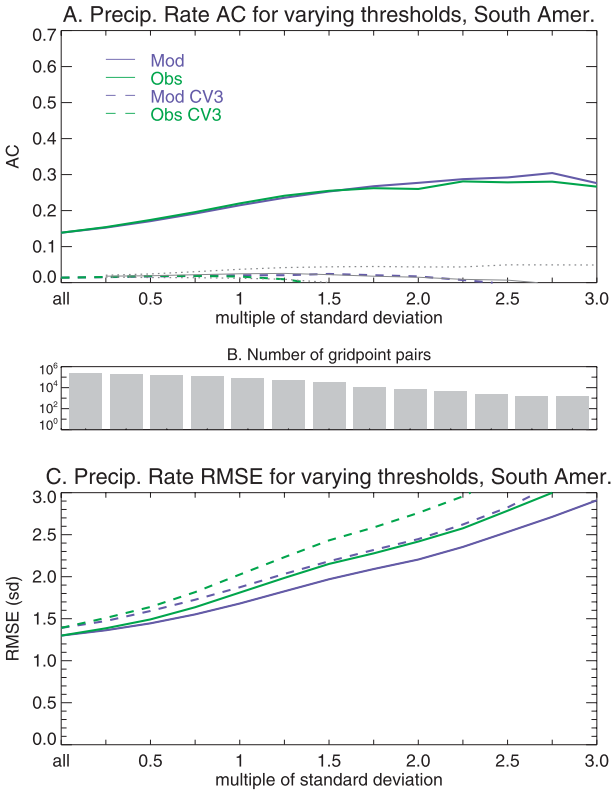

FIG. 5. As in Fig. 1, but for precipitation. ACs when cross validation is applied are near or below 0.

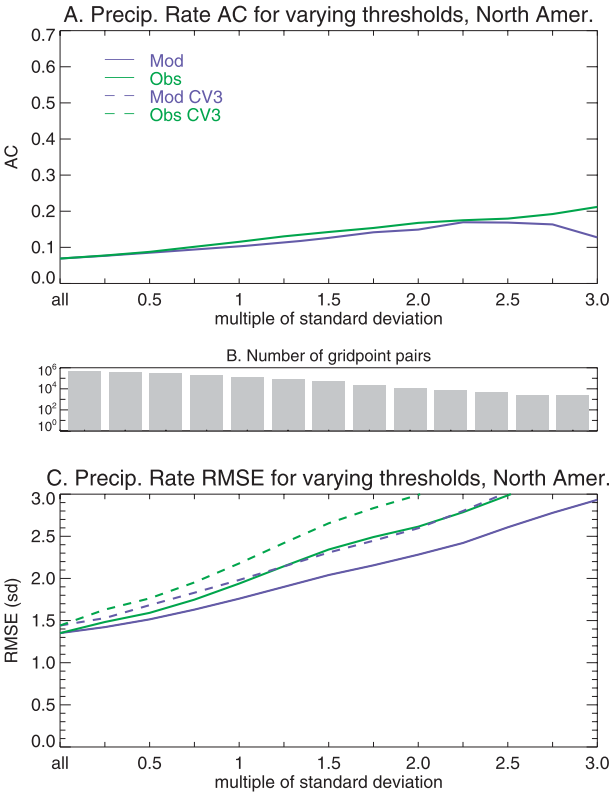

FIG. 6. As in Fig. 5, but for North America.

b. *Prediction skill: Precipitation*

Precipitation is notoriously difficult to predict given its sporadic, highly localized nature, even in monthly means. While anomaly correlations are expected to be lower for precipitation than for temperature, important regional and temporal patterns can still be detected, one hopes. Figures 5 and 6 (counterparts to Figs. 1 and 2 for temperature) show the area-averaged lead 1–3 anomaly correlations for precipitation rate over South and North America. We find that the rule of higher anomaly correlations when only extremes are examined (both forecast and observed) still holds, with ACs rising from around 0.13 to nearly 0.3 (Fig. 5) as the threshold increases. When cross validation is applied, ACs are at or below 0, so credible skill appears absent for precipitation, even for high thresholds. Over North America (Fig. 6) the AC is quite low, even in the non-cross-validated set, with ACs below 0.2 for extremes. RMSEs for both South (Fig. 5c) and North America (Fig. 6c) show higher error in the observed extremes subsample, but the slope of the increase in RMSE is still less than 45°. However, skill cannot be amplified by increasing the signal-to-noise ratio if there is no skill to begin with.

When we examine the skill of precipitation rate forecasts by target month in South America (Fig. 7), for the most part very limited skill is found. The all-cases scenario in the non-CV set (Fig. 7a) has ACs above 0.1 for most target months out to lead 2. Somewhat better scores are found in the non-CV set when only extremes are considered (Fig. 7b). Specifically, scores above 0.3 are found out to lead 2 in some target months in the first half of the year. In the cross-validated set (Fig. 7d), ACs after lead 1 fall off to below 0.1. Scores in North America (Fig. 8) are, unsurprisingly, lower yet, with little skill beyond lead 0. The qualified exception is extreme precipitation in the Northern Hemisphere winter (Fig. 8b), where somewhat higher ACs are found out to 2- or 3-month leads.

Precipitation rate has a predictably low hit rate for both wet and dry extremes (see Table 1), with the lowest HR over North America. Dry extremes are forecast substantially more often than they occur ($B > 1$), especially over North America, and positive extremes are underforecast. Distributions of monthly total precipitation are usually positively skewed, and so the definition of extreme as a ± 1.645 standard deviation may be cause for concern when examining this field. We tested the effect of this definition, which could result in too few dry

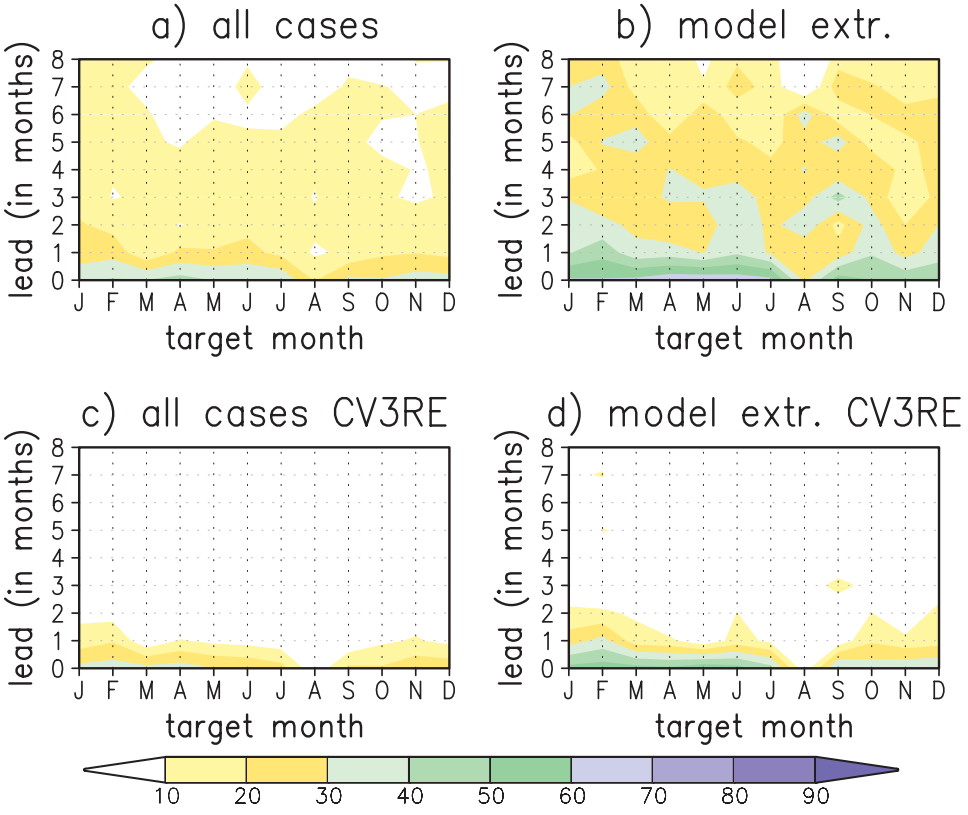

FIG. 7. As in Fig. 3, but for precipitation in South America.

extremes. First, before determining extremes we performed a power transform on the precipitation values, raising them to ¼ power, reducing the skewness of the distributions. A second test was to use the highest/lowest two of the 29 monthly values for each grid point. Both of these tests did result in more dry extremes going into the averages but did not substantially affect the anomaly correlations shown in Figs. 5, 6, 7, and 8.

c. *Prediction skill: Sea surface temperature*

Forecasts for sea surface temperature in the Niño-3.4 region (Fig. 9) start out with high ACs (around 0.78) for the all-cases scenario and increase as the sample is restricted to higher extremes, to greater than 0.9. Little difference is found between the results when the sample is filtered using predicted extremes or using observed extremes, and only very slightly lower scores are found when cross validation is applied. This is unsurprising, as cross validation has the greatest effect when scores are low (and possibly insignificant) to start with (van den Dool 2007). The bounds of the 1000 randomly selected subsamples (gray dotted lines) show that for smaller all-cases scenarios, the anomaly correlation, if anything, decreases. The Niño-3.4 region, bounded by 5°N–5°S, 190°–240°E, has fewer grid points (Fig. 9b) than the North or South American regions, and when the threshold for defining an extreme rises above 2.5, there are very few remaining grid points. The aggregated RMSE (Fig. 9c) for the grid points in this region is essentially flat, at approximately 0.7 for both predicted and observed extremes and for the non-CV and CV3RE sets. This is different from the behavior for precipitation rate and 2-m temperature (Figs. 1 and 5), where RMSE slowly increased. The argument that AC approximates the signal-to-noise ratio thus works even better to explain the AC rise under constant RMSE, as the threshold increases.

Another area of interest for short-term climate in the Americas is the Atlantic main development region (MDR), approximately 10° and 20°N, 275°–340°E, where the majority of Atlantic hurricanes have their genesis (Goldenberg and Shapiro 1996). Anomaly correlations in this region (Fig. 10a) are above 0.6 for the all-cases scenario, and rise to approximately 0.85 as the extremes-defining threshold is increased. The skill for the predicted and observed extremes is not substantially different, and cross validation with CV3RE results in only slightly lower ACs. The RMSE (Fig. 10c) shows a steady small increase, smaller than for T2m or precipitation but larger than the Niño-3.4 region, until a threshold of 2.25

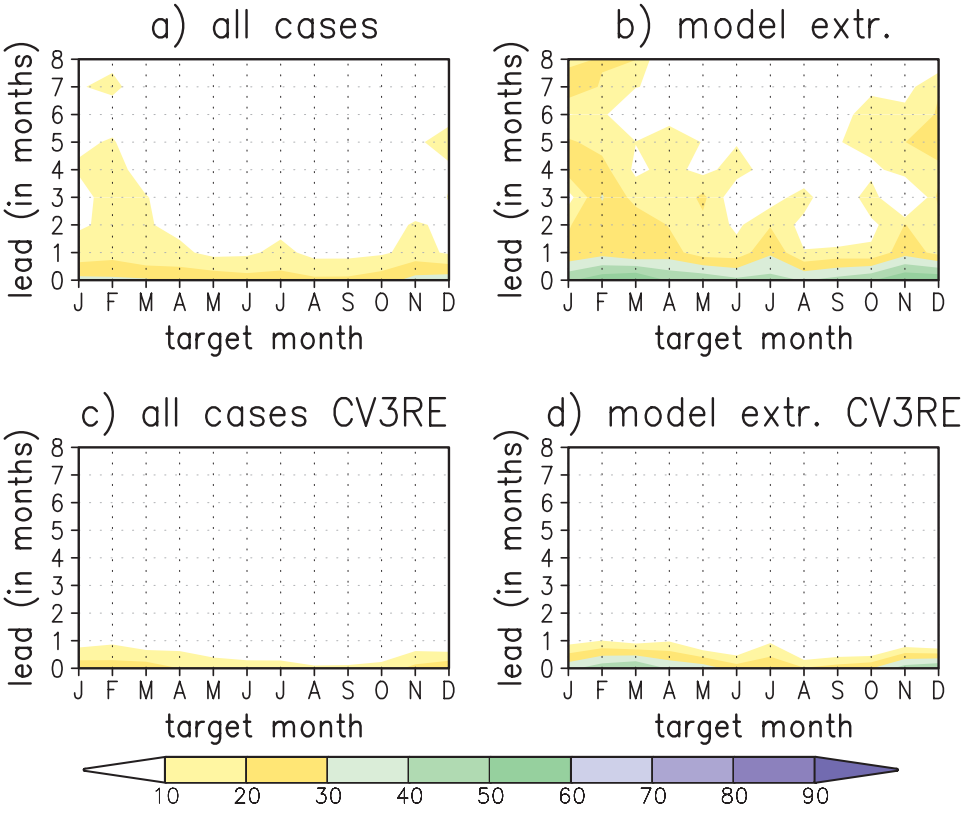

FIG. 8. As in Fig. 7, but for precipitation in North America.

standard deviations, at which point it begins to rise faster.

The skill of all-cases prediction of SST in the Niño-3.4 region (Fig. 11a) is high for most target months out to 8-month leads. The weakest scores are for the boreal spring, especially May, and for September, at 5-month and longer leads. Cross validation (Fig. 11c) leads to slightly lower scores at longer leads, and has almost no effect at shorter (less than 4 month) leads. The prediction of extremes (Figs. 11b,d) has very high scores, with the only ACs below 0.7 occurring at 7- to 8-month lead forecasts for July and August.

Anomaly correlations for prediction of SST in the MDR area are above 0.6 out to 2–3 month leads for all target months for the overall case, and high out to long leads for extremes (Fig. 12). Anomaly correlations are not greatly diminished by cross validation. The highest long-lead scores are for the boreal summer and autumn, and ACs for the prediction of extremes are above 0.7 for these target months, out to 8-month leads. SSTs in this region are important for hurricane forecasting, so high forecasting skill during this season is welcome, especially for extremes.

With reference to Table 1, SST extremes in the Niño-3.4 region have hit rates around 0.44, both for warm and cold extremes. The bias ratio indicates that warm SST extremes in this region are underforecast, while cold SST extremes are somewhat overforecast. Warm SST extremes in the Atlantic hurricane MDR have reasonably high HRs, but cold extremes are often missed in the forecast. CFSv2 forecasts approximately the right number of both warm and cold extremes ($B \sim 1$).

5. Predictability

As an assessment of (potential) predictability under perfect model assumptions, the $N - 1$ member ensemble mean was verified against the one remaining member. Thus, we are testing how effectively the model predicts itself, and therefore the limit of predictability, if the model is a replica of reality. No cross validation is necessary here, since the operating assumption is that the single member is from the same world as the ensemble mean. Here we employ both the CFSv1 and CFSv2; predictability ideally should not depend on the model and should be similar when assessed using either model. Figures 13–15 show the results of our predictability experiments in the lead-target format employed in earlier figures. Since the outcome of the predictability experiment varies slightly depending on which ensemble

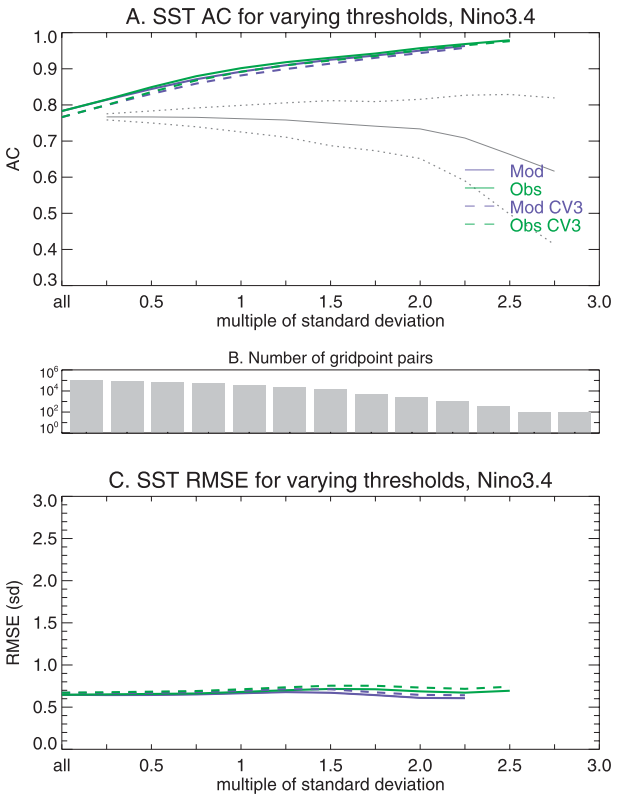

FIG. 9. As in Fig. 1, but for SST in the Niño-3.4 region, 5°S–5°N, 190°–240°E.

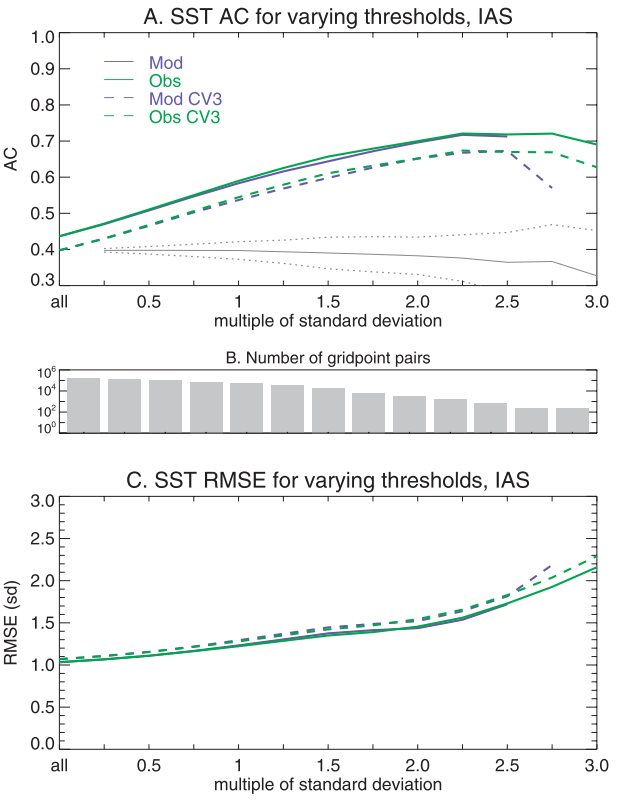

FIG. 10. As in Fig. 5, but for SST in the main development region (MDR), 10°–20°N, 275°–340°E.

member is used as the ''verifying'' values, predictability was calculated using each of the N members, and the results for all the ensemble members were averaged together.

Potential predictability was assessed for both the all-forecasts scenario and for extremes. For the extremes experiment, the anomaly correlation was calculated for all gridpoint pairs in the forecast where the single member (the one excluded from the ensemble mean) was in the extreme range [i.e., with a standardized anomaly greater (less) than 1.645 (−1.645)]. Results when the sample was filtered to gridpoint pairs with an extreme in the ensemble were also examined (not shown) and found to be similar to the former scenario.

The potential predictability of 2-m temperature in North America (Fig. 13, top) is highest in the boreal spring and summer. This is the case for both versions of CFS, although AC_p scores are slightly higher in CFSv2. Predictability of extremes is substantially higher than for all forecasts, and again, CFSv2 AC_ps are slightly higher. An AC_p value greater than 0.5 is found out to long leads during the boreal spring and summer. Comparing the predictability results to prediction skill (as seen in Fig. 4), we find the forecast skill is lower than its potential but is better for target months with higher predictability.

Potential predictability of 2-m temperature in South America (Fig. 13, bottom) shows a stronger dependence on season in CFSv2 than in CFSv1. The lowest AC_p scores in CFSv1 are at longer leads for target months of October–December. CFSv2 scores drop off starting in July, and are generally around 0.1 points lower than CFSv1 during the second half of the year. This result suggests the possibility that as models get better, and more complex, the predictability decreases. However, the predictability in the more complex model is likely a closer approximation of the true physical potential predictability. Again, as with North America, potential predictability is higher for extremes. CFSv2 shows the same strong seasonal dependence, with target months of January through April showing a higher potential than subsequent months. Interestingly, when we refer to the actual forecast skill for 2-m temperature in South America (Fig. 3) we find that while scores do drop after April, anomaly correlations for October, November, and December target months are somewhat higher.

The precipitation rate in North America (Fig. 14, top) has generally very low potential for prediction. Only for

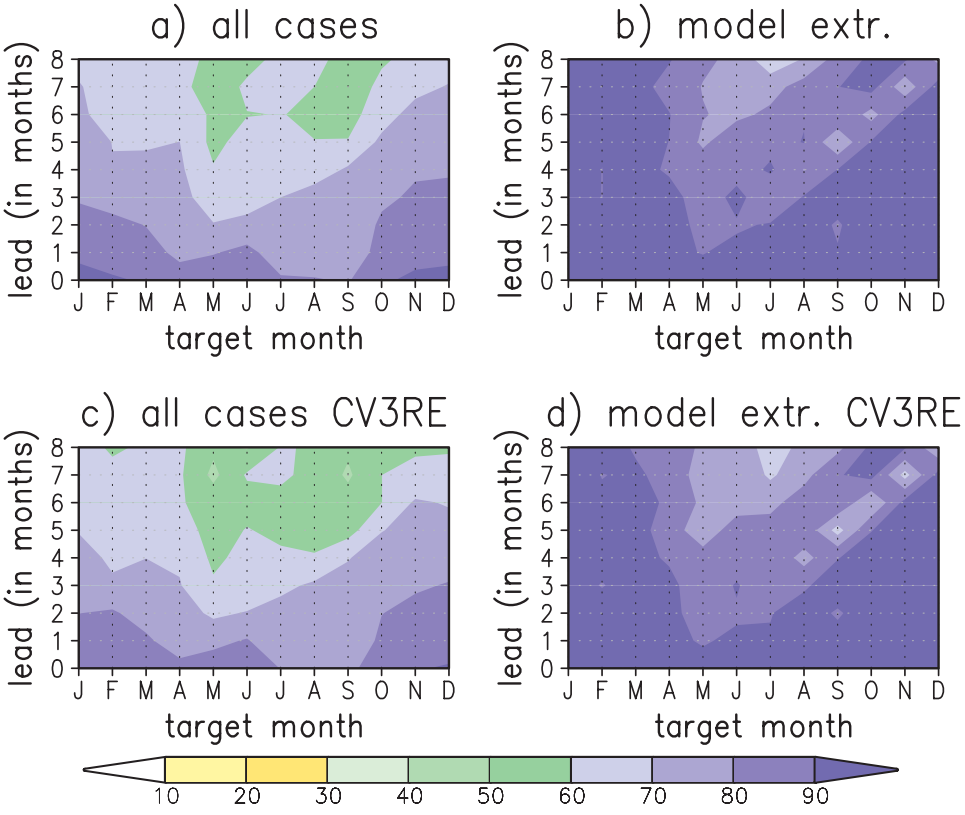

FIG. 11. As in Fig. 7, but for SST in the Niño-3.4 region.

the boreal winter and spring target months do AC_ps rise above 0.1; CFSv1 and CFSv2 reveal consistent results. It seems even under the best of circumstances (''perfect model'') the potential prediction skill for precipitation is very low. The potential predictability of extreme monthly mean precipitation is somewhat better during the winter target months in North America, however.

South American precipitation (Fig. 14, bottom) shows two bumps in potential predictability: for target months of January–March and July–October. Referring to the anomaly correlations for the forecast–observation pairs (Fig. 7), we find that while ACs during January–March are slightly higher than during the rest of the year, scores for July–October are the lowest of all target months, both for all cases and for extremes. Perhaps there is potential for improvement in the forecasts for precipitation rate during the austral winter.

Sea surface temperature potential predictability in the Niño-3.4 region (Fig. 15, top) is very high in general. In fact, AC_ps for extremes are above 0.9 at all leads and target months in CFSv1. However, CFSv2 shows AC_ps that are quite a bit lower than CFSv1 for June–August at the longer leads, both for all cases and for extremes. In view of Wu et al. (2009) this overstated predictability may be due the flawed tendency of CFSv1 to hang on to winter SST anomalies through the spring and summer. The high forecast–observation anomaly correlations of CFSv2 (Fig. 11) are approaching the potential predictability scores.

In the MDR region (Fig. 15, bottom), potential predictability scores are similar between CFS versions, with CFSv2 slightly higher at longer leads, especially for target months of March–May. Extremes in this region have high AC_ps out to long leads. The anomaly correlations found for the forecast skill (Fig. 12) do not show the highest scores during the periods of highest potential predictability.

6. Summary and discussion

In this study, we examined the model forecast skill of short-term climate extremes in 2-m temperature, precipitation, and sea surface temperature, using the CFS, in the region of the Americas. Twenty-nine years, 1982–2010, of the CFS version 2 monthly mean forecasts were available, for all 12 initial months, and we focused on extremes in the monthly mean. Metrics for skill include the anomaly correlation (AC) and root-mean-square error (RMSE). Since we have, for every grid point and time, two data points—one forecast and

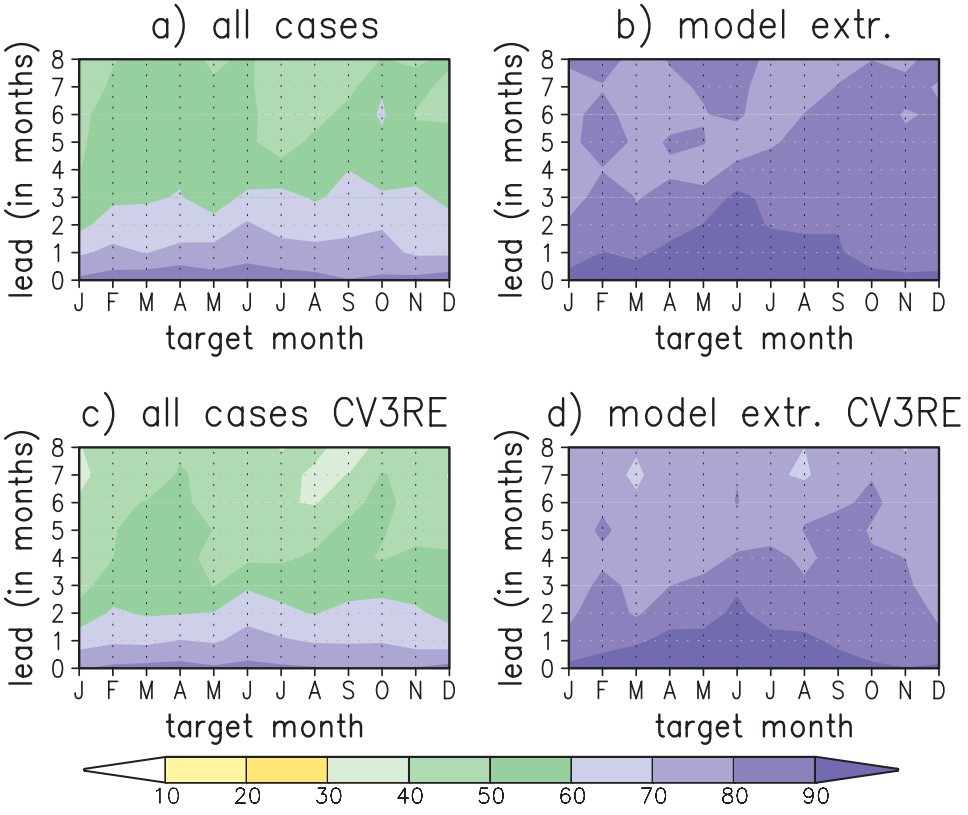

FIG. 12. As in Fig. 7, but for SST in the MDR region.

one observation—two different scenarios present themselves when we consider how to determine which cases to analyze. That is, we can include a case in the "extreme" subsample if the observation is extreme, or if the forecast is extreme. This allows us to study both how well the model captured extremes that occurred, and how well a forecast of an extreme "came true." The hit rate, false alarm rate, and bias in the forecasts were also studied through use of contingency tables.

Our investigation found that anomaly correlations for the short-term climate extremes subset are routinely higher than anomaly correlations for all forecasts, for both the "extreme observed" and the "extreme predicted" scenarios. This is not an artifact of the reduced sample size (using the ± 1.645 standard deviation threshold results in a subsample of about 10%). Further explanation of these higher scores is found in the RMSE, which can be considered the noise in a signal-to-noise ratio (of which the AC is one measure). While the RMSE does grow as the threshold defining an extreme is increased, it grows more slowly than the threshold, meaning that the signal grows despite increased noise. ACs for extremes are higher for all fields (T2m, precipitation, and SST). Scores for SST in the Niño-3.4 region are high for the all-cases scenario, and increase to nearly 1.00 for extremes; RMSE in this region is essentially the same for both all cases and extremes. It thus appears that the increase of RMSE with threshold is a function of inherent skill. For high skill, RMSE is constant with threshold, while with decreasing skill RMSE rises faster and faster as a function of threshold. Skill measures for the observed extreme and predicted extreme scenarios were generally very similar. While the higher skill for model-predicted extremes in North American 2-m temperature bears further investigation when more models are available for the analysis, the difference is likely small.

Cross validation can lead to lower anomaly correlations; ACs that are already low are reduced the most. The moderate (T2m) to low (precipitation) ACs are substantially lowered when analyzed under cross validation using CV3RE. In the case of precipitation, ACs under cross validation are near 0. Sea surface temperature ACs, which are relatively high, are only minimally affected by CV3RE, especially at short leads in the Niño-3.4 region.

Potential predictability of T2m under perfect-model assumptions finds the highest predictability for both North and South America is during the first half of the calendar year. Potential predictability in South America is higher than North America for both T2m and precipitation; precipitation AC_ps are very low in North

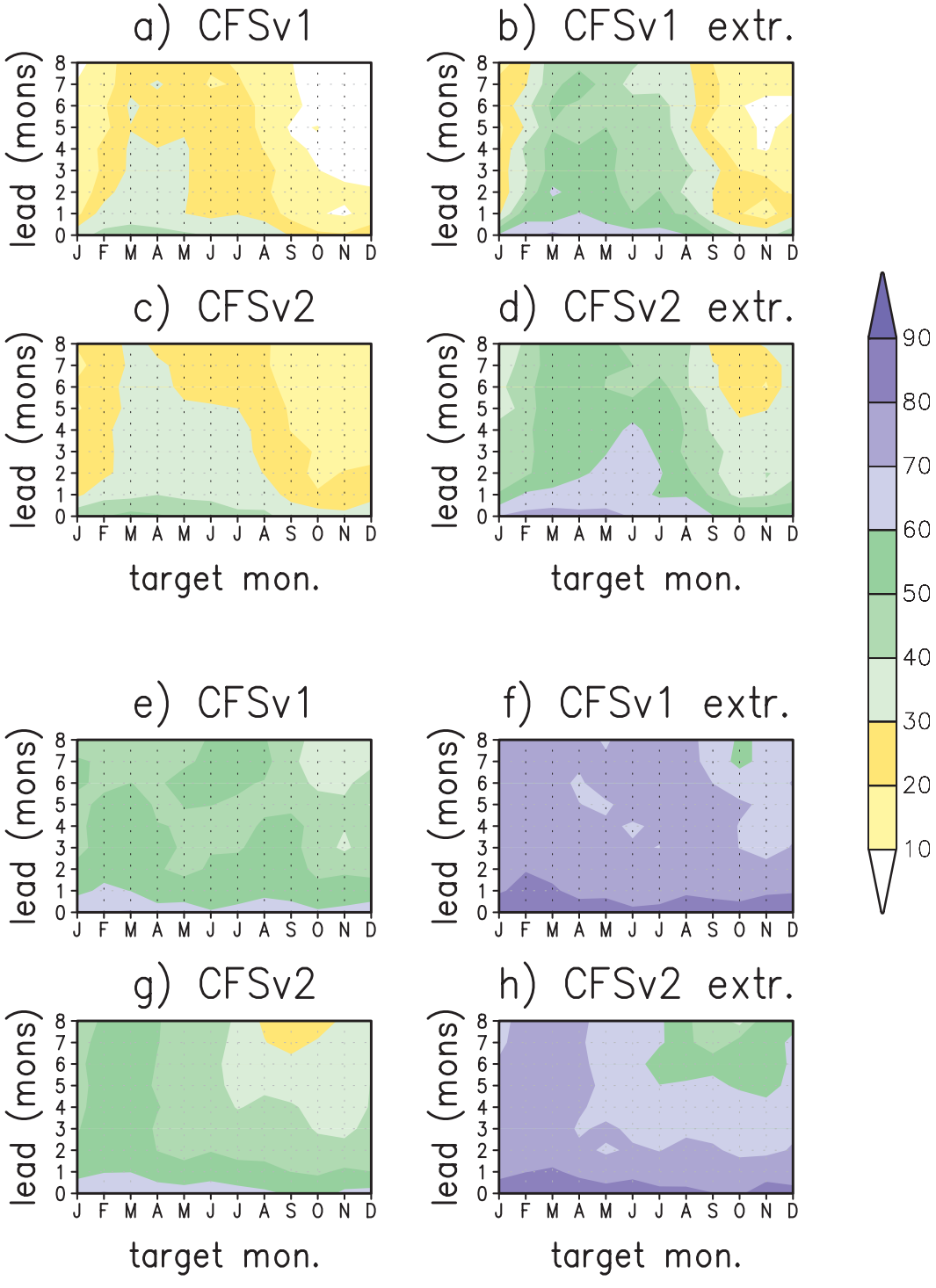

FIG. 13. Predictability, in the form of the anomaly correlation AC_p [see Eq. (4)], between a single member of the ensemble and the mean of the remaining members of the ensemble, as a function of target month, for 2-m temperature, for (top) North American and (bottom) South American regions, for (a),(e) CFsv1 all cases, (b),(f) CFSv2 extremes, (c),(g) CFSv2 all cases, and (d),(h) CFSv2 extremes. The AC_p values have been multiplied by 100.

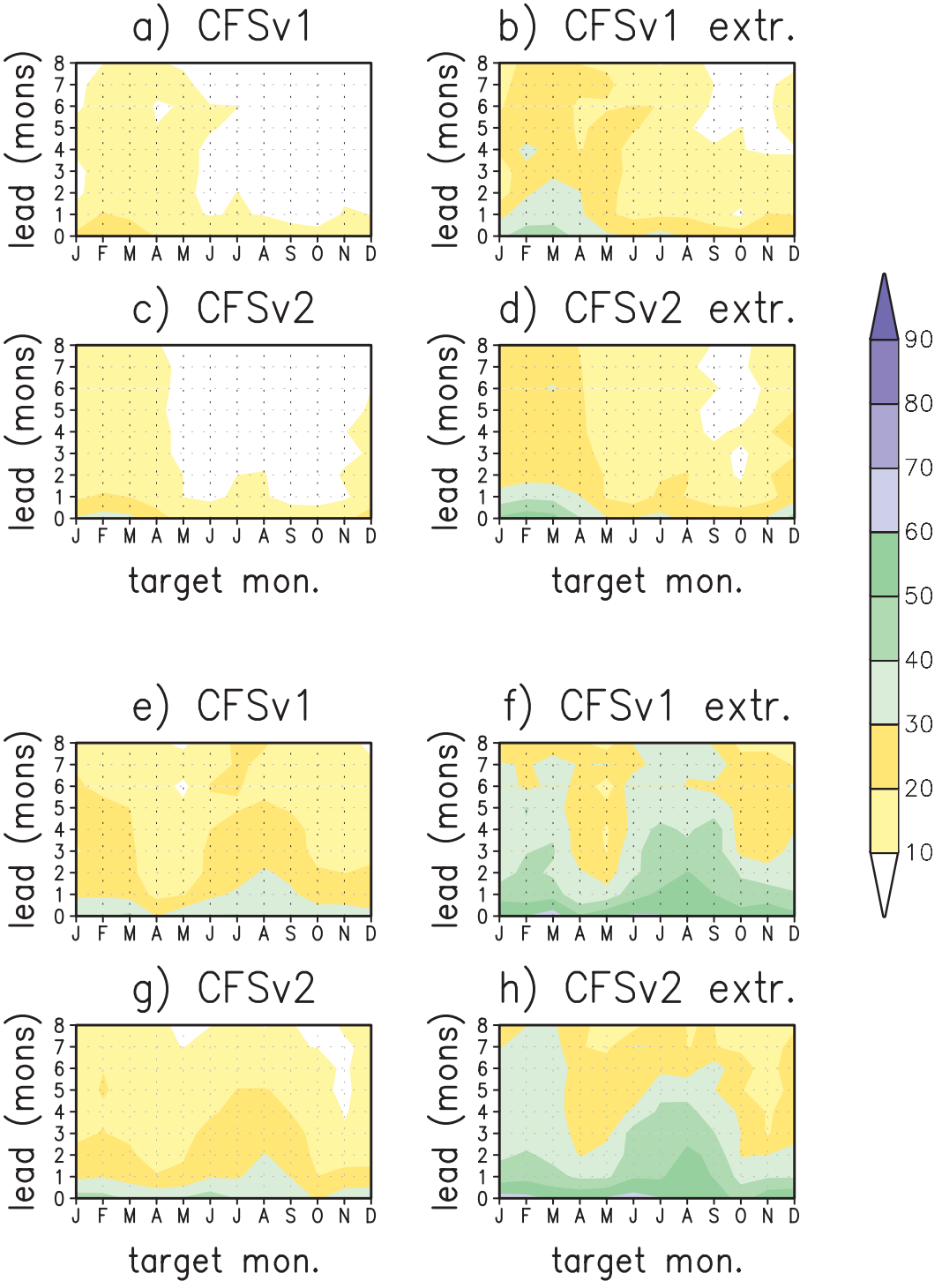

FIG. 14. As in Fig. 13, but for precipitation rate.

America overall. SST predictability is high in the Niño-3.4 region, and fairly high in the Atlantic hurricane main development region, with AC_ps above 0.5 at long leads for much of the hurricane season. Potential predictability of T2m, precipitation, and SST is generally slightly lower in CFSv2 than in CFSv1 (i.e., the estimate is model dependent, despite the perfect model assumption). One possible contributor to this effect is that as the model

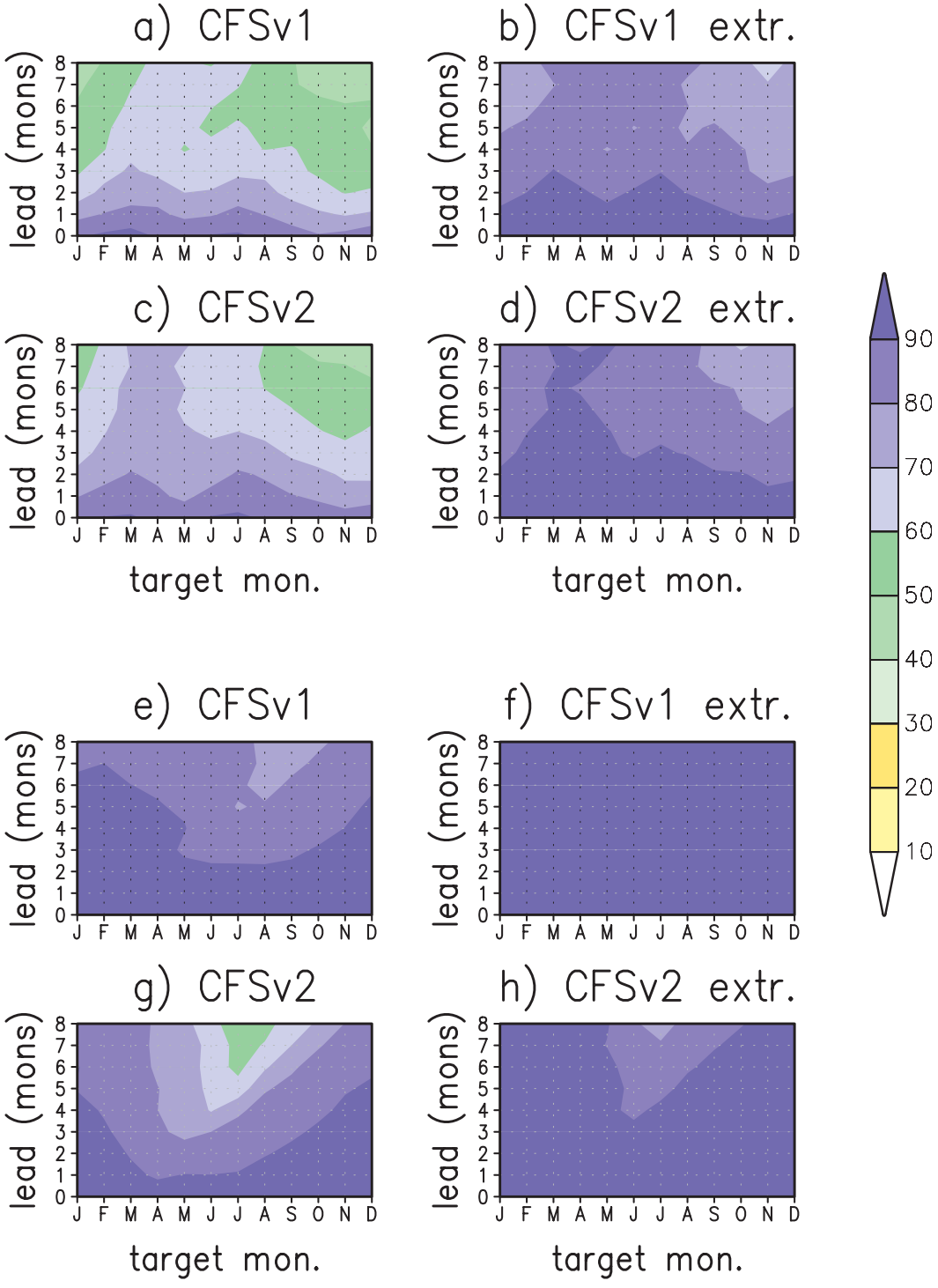

FIG. 15. Predictability, in the form of the anomaly correlation AC_p [see Eq. (4)], between a single member of the ensemble and the mean of the remaining members of the ensemble, as a function of target month (horizontal) and lead (vertical), for SST, for (top) Niño-3.4 region and (bottom) Atlantic hurricane MDR, for (a),(e) CFsv1 all cases, (b),(f) CFSv2 extremes, (c),(g) CFSv2 all cases, and (d),(h) CFSv2 extremes. The AC_p values have been multiplied by 100.

becomes more complex (and more accurate), predictability decreases as the model has a harder time predicting itself. Target months with higher potential predictability also have higher forecast skill, but forecast skill in general is lower than its potential.

Monthly means are generally a harder target to hit than 3-month seasonal means, with a lower signal-to-noise ratio in the former (Barnston 1994; Kumar et al. 2010). The higher forecast skill for 3-month means is probably due to the stronger predictive value of lower-frequency signals (Barnston 1994). This study focuses on extremes in the monthly mean, forecasts for which are of considerable practical importance to users. Tests for the skill of forecasts of extreme anomalies in the 3-month mean (not shown) show generally higher, less noisy anomaly correlation patterns, with similar areas of higher and lower scores.

As noted in section 3, there are many ways of defining an extreme. To test our definition, we examined ACs under several different extreme criteria. The first is the one discussed above: monthly mean standardized anomalies above/below ±1.645. A second technique involves taking the highest two and lowest two monthly anomalies for each grid point for the 27-yr dataset. This was applied for both the extreme observed and the extreme predicted scenarios. Finally, since the model comprises 24 ensemble members (28 in November), we assessed the results when extremes were defined by a number of the ensemble members with standardized anomalies of the same sign above–below ±1.645. For example, if five ensemble members are above 1.645, the ensemble mean is deemed to forecast a positive extreme. The results of these two latter analyses (not shown) found that while the sample size did vary with the definition, the area-averaged anomaly correlations within each field and location varied only slightly, with ACs by all definitions within 0.05 of each other. A fourth technique is the sliding definition as in Fig. 1, leaving the determination up to the reader.

Acknowledgments. The authors thank two anonymous reviewers whose attention and comments improved this paper. This work was supported by CPPA GC08-292, ''Predictability of drought and other near surface extreme events in the Americas.''

REFERENCES

Barnston, A. G., 1994: Linear statistical short-term climate predictive skill in the Northern Hemisphere. *J. Climate,* **7,** 1513–1564.

——, and H. M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate,* **6,** 963–977.

——, and S. J. Mason, 2011: Evaluation of IRI's seasonal climate forecasts for the extreme 15% tails. *Wea. Forecasting,* **26,** 545–554.

Chen, M., W. Wang, and A. Kumar, 2010: Prediction of monthly-mean temperature: The roles of atmospheric and land initial conditions and sea surface temperature. *J. Climate,* **23,** 717–725.

Compo, G. P., and P. D. Sardeshmukh, 2004: Storm track predictability on seasonal and decadal scales. *J. Climate,* **17,** 3701–3720.

Easterling, D. R., J. L. Evans, P. Ya. Groisman, T. R. Karl, K. E. Kunkel, and P. Ambenje, 2000a: Observed variability and trends in extreme climate events: A brief review. *Bull. Amer. Meteor. Soc.,* **81,** 417–425.

——, G. A. Meehl, C. Parmesan, S. A. Changnon, T. R. Karl, and L. O. Means, 2000b: Climate extremes: Observations, modeling, and impacts. *Science,* **289,** 2068–2074.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarplay, 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., and H. van den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.,* **113,** D01103, doi:10.1029/2007JD008470.

Goldenberg, S. B., and L. J. Shapiro, 1996: Physical mechanisms for the association of El Niño and West African rainfall with Atlantic major hurricane activity. *J. Climate,* **9,** 1169–1187.

Hurrell, J. W., 1995: Decadal trends in the North Atlantic oscillation: Regional temperatures and precipitation. *Science,* **269,** 676–679.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S. K. Yang, J. J. Hnilo, M. Fiorino, and G. L. Potter, 2002: NCEP–DOE AMIP-II Reanalysis (R-2). *Bull. Amer. Meteor. Soc.,* **83,** 1631–1643.

Kumar, A., M. Chen, and W. Wang, 2010: An analysis of prediction skill of monthly mean climate variability. *Climate Dyn.,* **37,** 1119–1131, doi:10.1007/s00382-010-0901-4.

Livezey, R. E., and M. M. Timofeyeva, 2008: The first decade of long-lead U.S. seasonal forecasts. *Bull. Amer. Meteor. Soc.,* **89,** 843–854, doi:10.1175/2008BAMS2488.1.

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus,* **34,** 505–513.

Madden, R. A., and P. R. Julian, 1972: Description of global-scale circulation cells in the tropics with a 40-50 day period. *J. Atmos. Sci.,* **29,** 1109–1123.

Mitchell, K., and Coauthors, 2004: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.,* **109,** D07S90, doi:10.1029/2003JD003823.

Nicholls, N., 1995: Long-term climate monitoring and extreme events. *Climatic Change,* **31,** 231–245.

NRC, 2010: *Assessment of Intraseasonal to Interannual Climate Prediction and Predictability*. The National Academies Press, 181 pp.

Rasmusson, E. M., and T. H. Carpenter, 1982: Variations in tropical sea surface temperature and surface wind fields associated with the Southern Oscillation/El Niño. *Mon. Wea. Rev.,* **110,** 354–384.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate,* **15,** 1609–1625.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate,* **19,** 3483–3517.

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

Savijarvi, H., 1995: Error growth in a large numerical forecast system. *Mon. Wea. Rev.,* **123,** 212–221.

Sheridan, S. C., and T. J. Dolney, 2003: Heat, mortality, and level of urbanization: Measuring vulnerability across Ohio, USA. *Climate Res.,* **24,** 255–265.

Smith, T. M., and R. E. Livezey, 1999: GCM systematic error correction and specification of the seasonal mean Pacific–North America region atmosphere from global SSTs. *J. Climate,* **12,** 273–288.

Stephenson, D. B., 2000: Use of the ''odds ratio'' for diagnosing forecast skill. *Wea. Forecasting,* **15,** 221–232.

van den Dool, H. M., 2007: *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, 215 pp.

——, and Z. Toth, 1991: Why do forecasts for ''near normal'' often fail? *Wea. Forecasting,* **6,** 76–85.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.,* **109,** 784–812.

Wang, W., M. Chen, and A. Kumar, 2010: An assessment of the CFS real-time seasonal forecasts. *Wea. Forecasting,* **25,** 950–969.

Wilks, D., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

Wolter, K., R. M. Dole, and C. A. Smith, 1999: Short-term climate extremes over the continental United States and ENSO. Part I: Seasonal temperatures. *J. Climate,* **12,** 3255–3272.

Wu, R., B. P. Kirtman, and H. van den Dool, 2009: An analysis of ENSO prediction skill in the CFS retrospective forecasts. *J. Climate,* **22,** 1801–1818.