Verification metrics should be chosen to answer specific questions regarding the quality of the forecast information. For example, they can identify where errors or biases exist in the forecasts to guide more effective use of them. The proposed questions address the accuracy in the forecast information and the representativeness of the forecast ensembles to indicate forecast uncertainty. Specifically, these questions are:

Q1: Do the initial conditions in the hindcasts lead to more accurate predictions of the climate?
Q2: Is the model's ensemble spread an appropriate representation of forecast uncertainty on average?
Q3: In the case that the forecast ensemble does offer information on overall forecast uncertainty, does the forecast-to-forecast variability of the ensemble spread carry meaningful information?

The first question, regarding whether the initial conditions provide a greater signal and thus greater accuracy in the predictions, can be addressed using deterministic metrics. We advocate the use of the mean squared skill score (MSSS) and its decomposition following Murphy (1988).

The mean squared skill score is based on the mean squared error (MSE):

$$MSE(f,x) = \frac{1}{n}\sum_{i=1}^{n}\left(f_i - x_i\right)^2, \tag{1}$$

where f_i *and* x_i are the forecast and observed values at time *i=1,n* years. The MSE is always positive unless the forecast identically matches the observations over time.
The general form of a skill score is written for some metric *A* as:

$$SS(f,r,x) = \left(A_f - A_r\right)/\left(A_p - A_r\right). \tag{2}$$

Here, the subscripts *f, r*, and *p*, represent the forecast system under test, the reference forecast, and the perfect forecast (i.e. MSE=0), respectively. The skill score is thus a function of the forecast, the reference, and the observations. The mean squared skill score for the common case where a climatological average for the period over which the hindcasts are being tested serves as the reference forecast can be written as:

$$MSSS(f,\bar{x},x) = r_{fx}^2 - \left[r_{fx} - \left(s_f/s_x\right)\right]^2 - \left[\left(\bar{f} - \bar{x}\right)/s_x\right]^2 \tag{3}$$

This summary metric actually constitutes a set of metrics: (1) the square of the correlation coefficient, (2) the conditional bias of the forecast, as indicated by the difference [squared] from unity between the slope of the regression between the forecast and observations, and (3) the unconditional bias in the forecast,

represented by the difference [squared] between the forecast and observed climatology. A positive value for the MSSS indicates that the forecast system is more accurate than the use of a climatological average. In the case of the initialized hindcast experiments, a more appropriate reference forecast is the set of uninitialized hindcasts. In this case, the reference forecast changes with the forecast under test and is likely to have similar biases. The MSSS in this case takes the form:

$$MSSS(f,r,x) = \left\{ r_{fx}^2 - \left[r_{fx} - \left(s_f / s_x \right) \right]^2 - \left[\left(\bar{f} - \bar{x} \right) / s_x \right]^2 - r_{rx}^2 + \left[r_{rx} - \left(s_r / s_x \right) \right]^2 + \left[\left(\bar{r} - \bar{x} \right) / s_x \right]^2 \right\}$$
$$/ \left\{ 1 - r_{rx}^2 + \left[r_{rx} - \left(s_r / s_x \right) \right]^2 + \left[\left(\bar{r} - \bar{x} \right) / s_x \right]^2 \right\}. \qquad (4)$$

Additional terms appear in the MSSS similar to (1)-(3) above, but considering these reference forecasts. The MSSS here represents the fractional improvement, or degradation, of the initialized hindcasts over the uninitialized ones. By multiplying the MSSS of Eqn. 4 by 100, one quantifies the percentage improvement in accuracy of the initialized hindcasts compared to the uninitialized ones. An example of the MSSS (Fig. 1) for year 1 annual-mean precipitation hindcasts from the DePreSys system initialized at the end of years 1960-2005 shows that unconditional and conditional biases exist in the model, which are quite similar between forecast design. The correlation coefficient (Fig 1), which is merely the linear association between the forecast mean and the observations (e.g. Murphy 1988), gives essential a measure of potential skill in that the translation between the forecast value and the observed value must consider the biases inherent in the forecasts.

Even as a measure of potential skill the correlation as implemented above, which is the Pearson's correlation (e.g. Fig 2a) , is known to be sensitive to outliers. Thus it may be informative to also consider alternate correlation measures, such as the Spearman's R Ranked Correlation (Fig 2b) or the Kendall's Tau Rank Correlation (Fig 2c). Thus the wettest or hottest forecasts are expected to lead to the wettest or hottest observations, even if the linear scaling may not be preserved.

The Spearman's R Rank Correlation coefficient is essentially the correlation coefficient of the time series of the ranked data. The observed and forecast vectors X and F are transformed into their rank in ascending order, Rx_i and Rf_i. When there are no ties in the ranks, i.e. when there are no identical values in either of the timeseries, then the correlation can be calculated as:

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} (Rx_i - Rf_i)^2}{n(n^2 - 1)} \quad , \qquad (5)$$

Otherwise, when there are ties in the ranked data, the Pearson's correlation is calculated using the ranked data, where an averaged rank is given to the repeated values. The Kendall's Tau Rank Correlation considers all possible comparisons of forecast and observation pairs, i.e. (x_i,f_i) vs (x_j,f_j). They are considered concordant if the rank of both elements agree, that is: $x_i > x_j$ and $f_i > f_j$, or if $x_i > x_j$ and $f_i > f_j$.

Otherwise they are considered discordant. The Kendall Tau coefficient is then calculated as:

$$\tau = \frac{(\#\ \text{concordant pairs}) \ - \ (\#\ \text{discordant pairs})}{1/2n(n-1)} . \tag{6}$$

In addition to establishing the level of accuracy in the ensemble mean forecast, one is often interested in quantifying the range of possibilities or uncertainty about that forecast value. This question requires the use of probabilistic metrics. The purpose of the probabilistic metric here is not to ascertain skill of the forecast, however, as that would be largely redundant information to what the deterministic metrics yielded. Here we pose the question of whether the ensemble spread in the forecast is, on average, adequate to represent the forecast uncertainty. Again, a skill score is used to determine the probabilistic quality of the forecast spread relative to some reference approach. The measure of probabilistic quality is the continuous ranked probability score (CRPS). The ranked probability score is commonly used to assess probabilistic forecasts (e.g. Barnston et al. 2010), but is typically used with categorical forecasts. Since the changing background climate subverts the usefulness of categorical forecasts, we wish to cast the forecasts in terms of a continuous, quantitative, analytical distribution with a mean and standard deviation determined from the forecast ensemble, although clearly both of these parameters are subject to substantial sampling errors with the small nominal ensemble sizes requested for CMIP5.

By definition, the CRPS is:

$$CRPS(f_i, x_i) = \int_{-\infty}^{+\infty} \left(G(f_i) - H(x_i) \right)^2 dy \tag{7}$$

where G and H represent the cumulative distribution functions of the forecast and the observations, respectively. Thus the CRPS is very much like the mean squared error, but in probability space. If one had infinitely confident forecasts (i.e. probability of 100%) for the observed outcome in every case, the CRPS would be 0.

In this case, where *x*, represents the observations, the cumulative function *H* is the Heavyside function. If the predictive distribution is a Gaussian with mean *f* and variance σ^2, then it follows that (Gneiting and Raftery, 2007):

$$CRPS(N(f_i, \sigma_f^2), x_i) = \sigma_f \left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{x_i - f_i}{\sigma_f} \right) - \frac{x_i - f_i}{\sigma_f} \left(2\phi\left(\frac{x_i - f_i}{\sigma_f} \right) - 1 \right) \right] \tag{8}$$

where φ and ϕ represent the probability distribution function (pdf) and cumulative distribution function (cdf) of a standard Gaussian variable. Given the $CRPS_f$ for the

forecast distribution, and the $CRPS_r$ for the reference distribution, the corresponding skill score can be defined as:

$$CRPSS = 1 - \frac{\sum_{i=1}^{n} CRPS_{f_i}}{\sum_{i=1}^{n} CRPS_{r_i}} \tag{9}$$

The "forecast" distribution is assumed Gaussian, with the mean given by the ensemble mean and the variance given by the average ensemble variance (i.e. averaged over all hindcasts). Since we are only testing the uncertainty in the forecasts, the mean of the distribution is the same for both the forecast under test and the reference forecast (i.e. $r_i = F_i$). The "reference" distribution has a variance given by the standard error variance of the hindcasts' ensemble mean compared to the observations.

For a given lead time or average of lead times, the observations and the forecast can be represented as:

$$\begin{aligned} &x_i, i = 1,\ldots,n \\ &F_i^e, i = 1,\ldots,n \;; e = 1,\ldots,N \end{aligned} \tag{10}$$

such that the average over all ensemble members, *e*, in any particular year, *i*, yields the ensemble mean value, F_i. Note that we have changed notation from *f* used in the section on deterministic metrics, which refers to the raw forecast data, to *F* here, which implies some degree of bias correction. For the results shown, the forecasts have been bias corrected appropriate to their experimental design, according to the recommended guidelines of WCRP (WCRP 2011). However, there may still exist conditional as well as unconditional biases in the forecast information.

Probabilistic Case 1 (Q2): Is the model's ensemble spread an appropriate representation of forecast uncertainty on average?

For the forecast distribution, the variance of the "forecast" distribution would be:

$$\overline{\sigma}_f = \frac{\sum_{i=1}^{n} \sigma_{f_i}}{n} \tag{11}$$

and for the reference distribution:

$$\overline{\sigma}_r = \frac{\sum_{i=1}^{n} \sqrt{\left(F_i - x_i\right)^2}}{n} \tag{12}$$

It should be noted that if biases remain in the forecasts, the standard error in the linear regression between the forecast mean and the observations, $\overline{\sigma}_r$, may actually be larger than the climatological variance of the observations. This is another reason that forecast data should be used judiciously.

According to our example, the forecast uncertainty using the standard error is comparable to that using the average spread of the forecast members in many locations (Fig 3). In some places, however, the standard error leads to better reliability. These cases likely reflect an under-dispersion of the ensemble members, a situation that is common in dynamical seasonal forecasts.

Probabilistic Case 2 (Q3): Does the forecast-to-forecast variability of the ensemble spread carry meaningful information?

Even if the average forecast spread is not a better representation of forecast uncertainty than the standard error of the forecast mean, there may still be information contained in the variability of the forecast spread. Such a situation could be seen if tighter, more confident, forecasts were also more accurate.

Again, the mean of the "forecast" and "reference" distributions has a mean given by forecast ensemble mean, F_i. The uncertainty for the forecast under test, σ_f, changes each year based on the spread of the ensemble members for each particular forecast. The "reference" distribution has the spread given by the average ensemble spread, which was defined above as $\overline{\sigma}_f$.

We note that there are some locations where the forecast-to-forecast ensemble spread improves upon the use of an average spread (Fig. 4). Given the rare and somewhat noisy appearance of these areas, we would conclude rather that the difference may be related to sampling errors, and it is better to conclude that the two approaches are comparable. On the other hand since the map is dominantly indicating negative values, the message is that the use of an average ensemble spread to indicate forecast uncertainty is the more prudent approach with this model.

References:

Barnston, A.G., S. Li, S.J. Mason, D.G. DeWitt, L. Goddard, X. Gong, 2010. Verification of the First 11 Years of IRI's Seasonal Climate Forecasts. *J. App. Met. and Clim.*, 49: 493-520.

Gneiting, T. and A.E.Raftery, 2007. Strictly Proper Scoring Rules, Prediction, and Estimation, *J. Amer. Stat. Assoc.*, **102**, 359-378

Murphy, A.H., 1988. Skill scores based on the Mean Square Error and their relationships to the correlation coefficient. *J. Clim.*, **116**, 2417-2424.

MSSS : Precipitation, Year 2-9

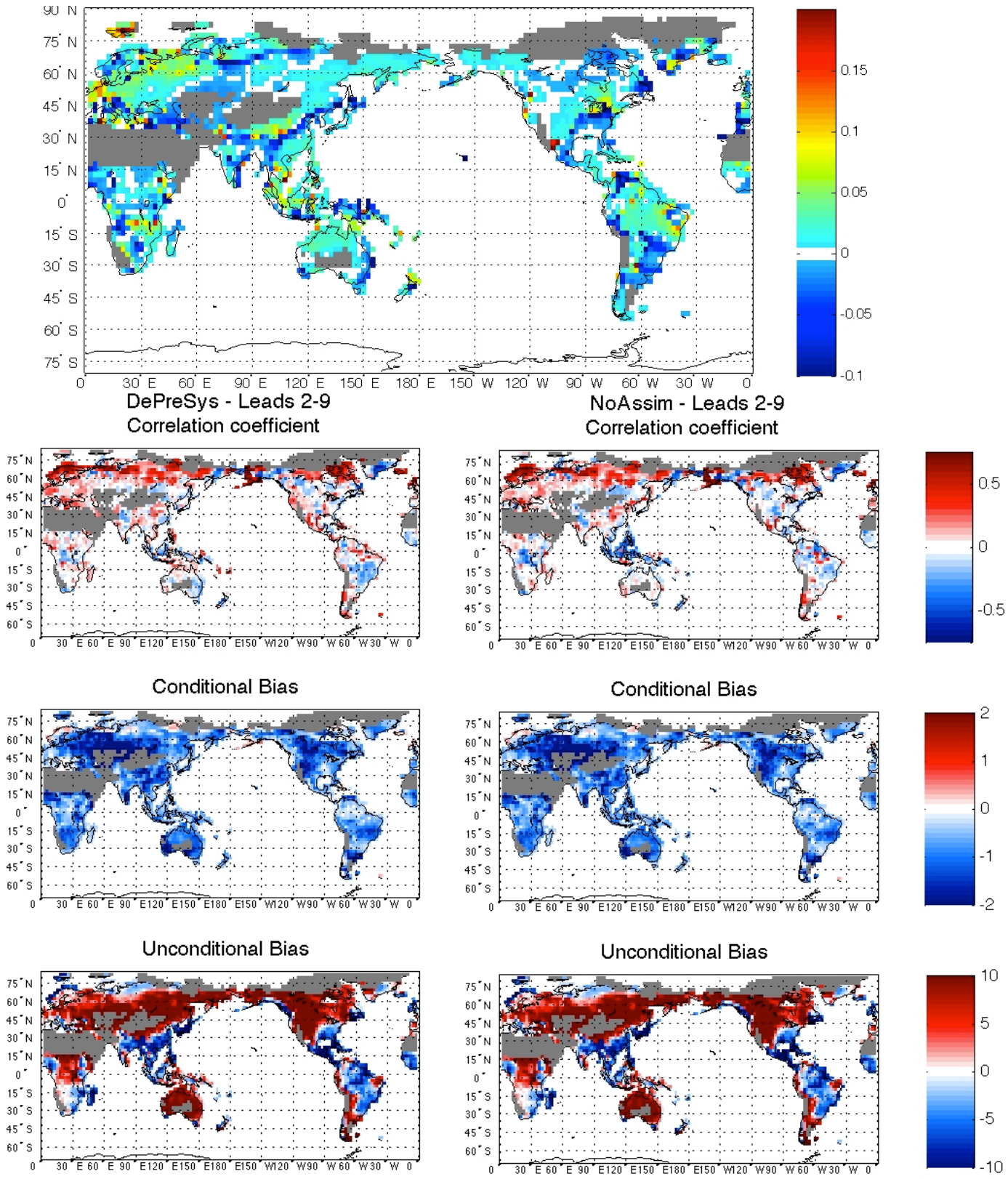

Figure 1. Mean squared skill score (MSSS) comparing the initialized decadal hindcasts ("forecast") against the uninitialized hindcasts ("reference") for annual mean precipitation of years 2-9, averaged, for forecasts initialized near the end of the years 1960-2005. Below is the decomposition of terms in the MSSS for both the "forecast" (left) and "reference" (right) data.

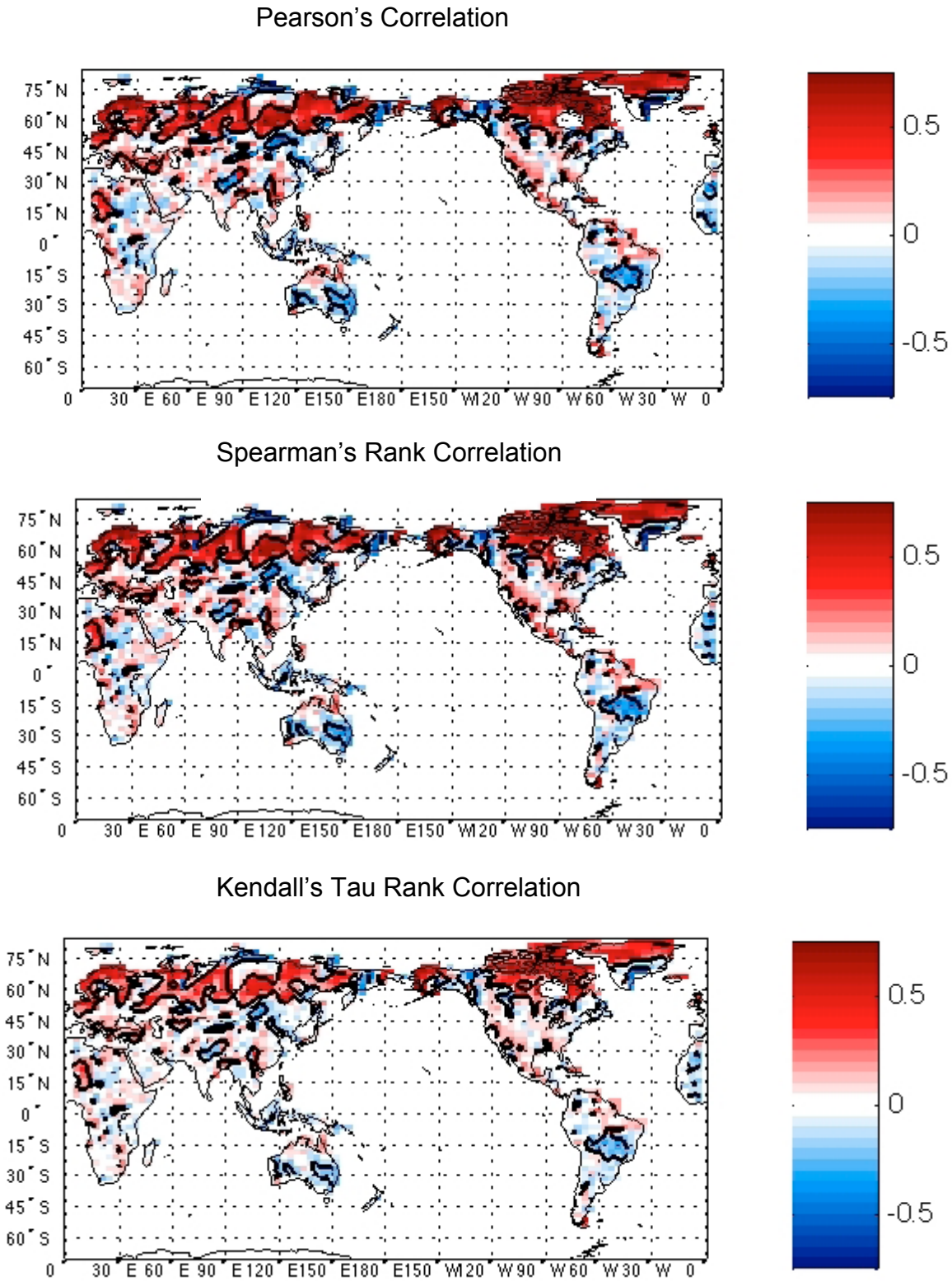

Figure 2. Different correlation metrics applied to initialized decadal hindcasts for annual mean precipitation of years 2-9, averaged, for forecasts initialized near the end of the years 1960-2005.

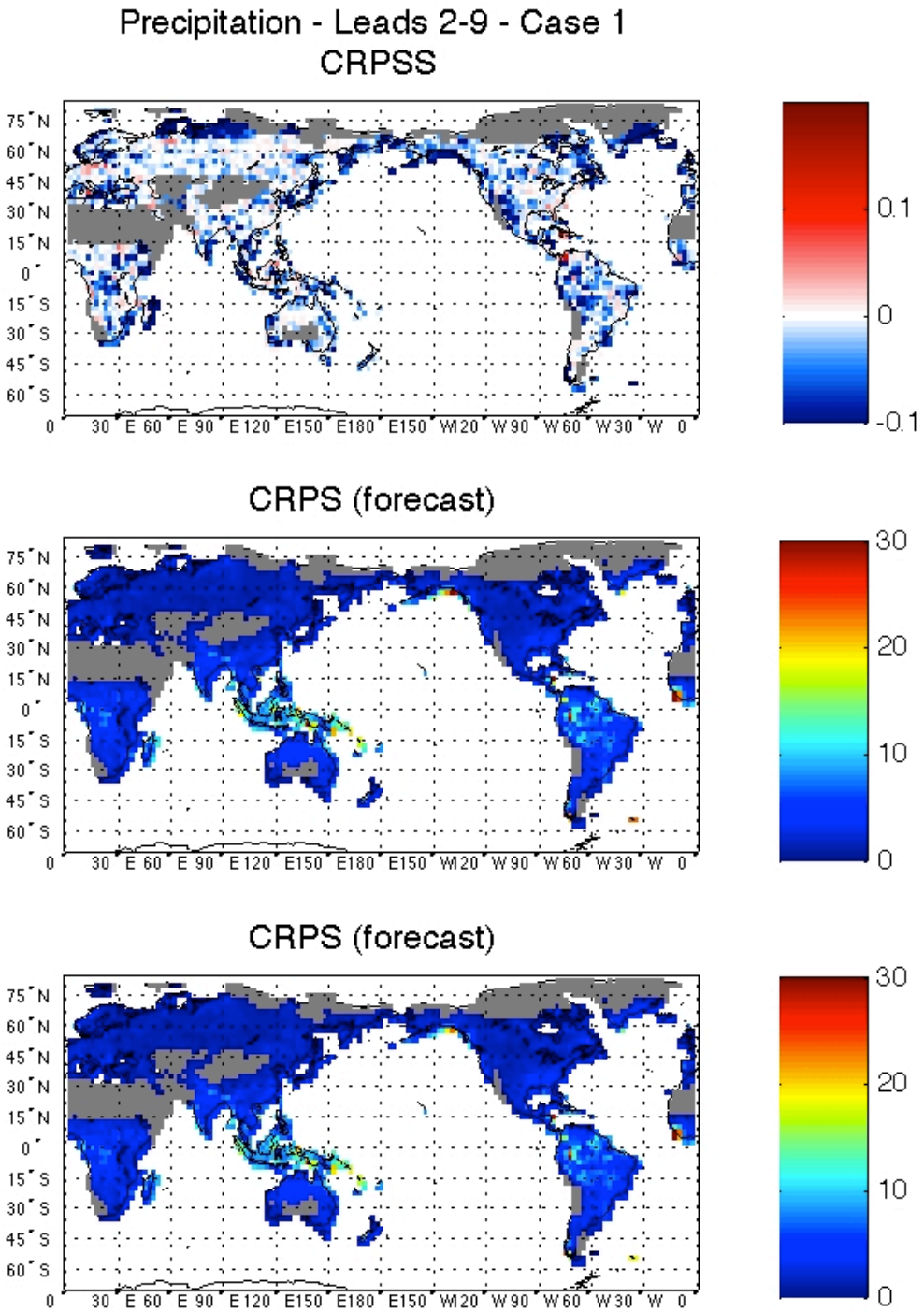

Figure 3. The continuous ranked probability skill score (CRPSS, top) comparing the continuous ranked probability score (CRPS) between the forecast distribution using the average ensemble spread of the forecasts (middle) against distributions that use the standard error of the forecast mean (bottom) to represent uncertainty. The forecast variable is annual mean precipitation from years 2-9, averaged.

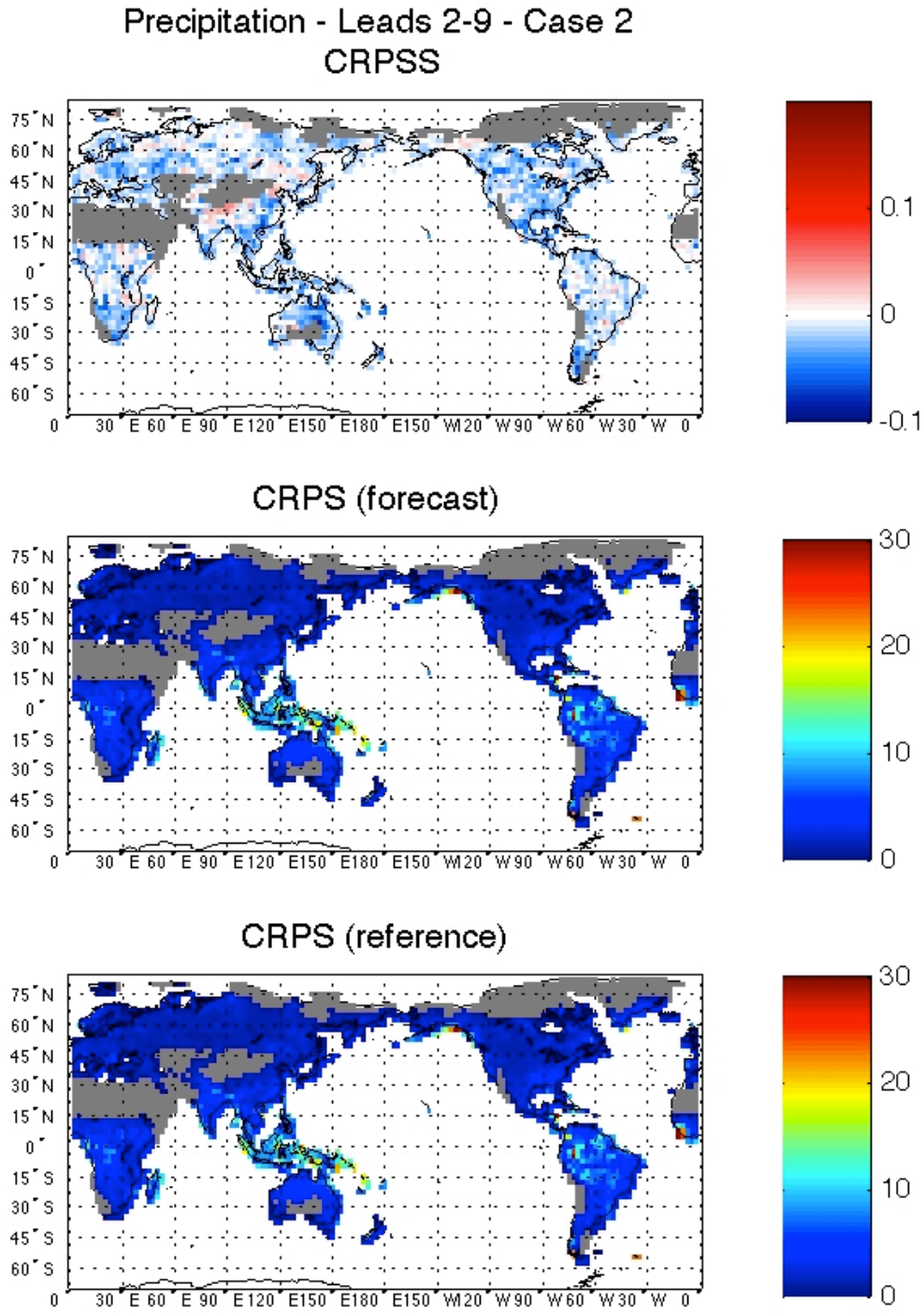

Figure 4. The continuous ranked probability skill score (CRPSS, top) comparing the continuous ranked probability score (CRPS) between the forecast distributions using the forecast-specific ensemble spread (middle) against distributions that use the average ensemble spread of the forecasts (bottom) to represent uncertainty. The forecast variable is annual mean precipitation from years 2-9, averaged.