

# Prioritizing Data for Improving the Multidecadal Predictive Capability of Atmospheric Models

STEPHEN S. LEROY

*Harvard School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts*

GIANLUCA REDAELLI AND BARBARA GRASSI

*CETEMPS/Department of Physical and Chemical Science, University of L'Aquila, Coppito-L'Aquila, Italy*

(Manuscript received 25 June 2014, in final form 6 October 2014)

## ABSTRACT

The prioritization accorded to observation types currently being considered for a space-based climate observing system is extended from a previous study. Hindcast averages and trends from 1970 through 2005 of longitude–latitude maps of 200-hPa geopotential height and of net downward shortwave and longwave radiation at the top of the atmosphere are investigated as relevant tests of climate models for predicting multidecadal surface air temperature change. To discover the strongest tests of climate models, Bayes's theorem is applied to the output provided by phase 5 of the Coupled Model Intercomparison, and correlations of hindcasts and multidecadal climate prediction are used to rank the observation types and long-term averages versus long-term trends. Spatial patterns in data are shown to contain more information for improving climate prediction than do global averages of data, but no statistically significant test is found by considering select locations on the globe. Eigenmodes of intermodel differences in hindcasts may likely serve as tests of climate models that can improve interdecadal climate prediction, in particular the rate of Arctic tropospheric expansion, which is measurable by Earth radio occultation.

## 1. Introduction

The work presented herein incorporates the approach to prioritizing data types presented in [Huang et al. \(2011, hereinafter H11\)](#). H11 sought to modify the projection of a 50-yr trend in global average temperature using varying lengths of trends in several observation types, both in situ and remotely sensed. The ensemble of models used for climate projection was the model output from the World Climate Research Programme (WCRP)'s phase 3 of the Coupled Model Intercomparison Project (CMIP3) multimodel dataset. Increasing the length of the data record reduced the range of climate prediction and improved its accuracy. Remotely sensed observation types, especially high-spectral-resolution thermal

infrared spectra, reduced uncertainty in climate projection better than did the in situ observation types. The conclusiveness of the study was limited, though, by the small size of the climate model ensemble of opportunity used. Using a larger ensemble of climate models, either an ensemble of opportunity or a perturbed physics ensemble, is needed for follow-on research.

Here we expand on [H11](#) by considering spatial patterns in the prospective observation types and both their long-term averages and trends. We also use the multimodel dataset from WCRP's phase 5 of the Coupled Model Intercomparison Project (CMIP5; [Taylor et al. 2012](#)) in place of the CMIP3 ensemble. By considering resolved spatial patterns in prospective observation types, it is possible that important information about climate useful for climate prediction can be revealed that is obscured in a global average. For example, the global average trend in reflected shortwave radiation is caused by a combination of decreasing albedo in the Arctic, decreasing snow and ice in midlatitudes, and changes in cloud fraction especially in low latitudes

---

 Denotes Open Access content.

---

*Corresponding author address:* Stephen Leroy, Anderson Group, 12 Oxford St., Link Building, Cambridge, MA 02138.  
E-mail: leroy@huarp.harvard.edu

DOI: 10.1175/JCLI-D-14-00444.1

(Feldman et al. 2011, 2013); hence, spatially resolving trends in reflected shortwave radiation may help in attributing trends in reflected shortwave to well-known feedbacks in the climate system. Even though it has been customary to consider long-term trends in observations as relevant to testing climate models according to their ability to predict long-term change, it has never been proven that long-term trends are more relevant a metric than long-term averages or the annual cycle. Already there are suggestions that testing climate models against long-term averages of observations can reduce climate prediction uncertainty (Geoffroy et al. 2012) as can an annual cycle (Hall and Qu 2006). Here we consider both long-term trends in observation types and long-term averages in those observation types as prospective metrics in testing climate models. Because we are concerned with the prioritization of climate monitoring observation types and the annual cycle does not necessarily require long time series of data, we do not address the annual cycle as a metric in testing climate models in this work.

In the second section we give a brief description of the method of testing climate models previously presented in H11. The third section explores the spatial signatures in long-term averages and trends of various observable types that can inform multidecadal global average temperature change, and the fourth section presents an eigen-decomposition of those spatial signatures. The fifth summarizes our results and presents conclusions.

## 2. Approach

We use methods of ensemble climate prediction to prioritize observation types for climate monitoring. Ensemble climate prediction results in a probability density function of a future climate based on predictive runs of a wide variety of climate models subjected to a prescribed set of external boundary conditions. Each of the member models of the ensemble has been adapted or modified to match some previously obtained data that its custodian had determined mandatory to reproduce. While different modeling groups can choose different data for the test, it is already common that they test against multidecadal trends in global average surface air temperature and best estimates of the present-day top-of-atmosphere radiative imbalance. Many other datasets, some of questionable quality, are available for use in testing, but it is not obvious which should be considered next. Prioritization of observation types is intended to answer this question.

One way to prioritize observation types is by weighting models according to how well they match observations.

First, a probability density function of climate projection is constructed using the forecasts of all the member models of an ensemble. Next, a weight is determined for each member model according to how well its hindcasts of data compare to observed data (with an allowance for uncertainty due to internal variability). The weights are “Bayes factors” (Min and Hense 2006a,b). Finally, an updated probability density function of climate projection is constructed by applying the models’ weight factors to their forecasts. The relative priority of a candidate observation type is determined by the ratio of the width of the unweighted probability density function to the width of the weighted probability density function.

Another method to prioritize data types is by Bayes’s theorem (Sivia 2006) applied to ensemble prediction. In this method, a large ensemble of climate models is used to populate a joint probability density function of hindcast observation and predicted climate change, and that distribution is modified using a likelihood function in which the data are conditioned on the hindcast (i.e., how likely that the data were obtained if a specific hindcast was true of the climate system?). Application of Bayes’s theorem enables the generation of a posterior probability density function of ensemble prediction after incorporation of data. The ratio of the prior prediction uncertainty and the modified, posterior prediction uncertainty is used to prioritize the observation types and metrics.

It can be shown that, in the limit of an ensemble with an infinite number of members, the two methods of prioritizing observation types presented above are the same. The former is modification of a climate ensemble by conditioning models on data, which is a simple way of describing Bayes’s theorem. We choose to work with the latter method because it permits us to circumvent the lack of an initial condition ensemble for each model and because it is in common use, having been applied in a number of previous studies attempting to infer climate sensitivity from model ensembles modified by data (Forest et al. 2002; Sexton et al. 2012).

We briefly describe the mathematics of H11, a simple application of Bayes’s theorem that approximates the joint probability density function of hindcast and climate prediction by the CMIP3 ensemble of climate models as a multidimensional Gaussian. We make the same approximation in this research. The hindcast observation by climate models is  $\mathbf{x}$  and the climate prediction is  $\mathbf{y}$ ; the ensemble mean hindcast and the ensemble mean climate prediction are  $\alpha_{\mathbf{x}}$  and  $\alpha_{\mathbf{y}}$ , respectively. Across the ensemble, the covariance in the spread of hindcasts is  $\Sigma_{\mathbf{xx}}$ , the covariance in the spread of climate predictions is  $\Sigma_{\mathbf{yy}}$ , and the covariance between them is  $\Sigma_{\mathbf{xy}} = \Sigma'_{\mathbf{yx}}$ .

(The prime indicates a matrix transpose.) The prior joint probability density function  $P(\mathbf{x}, \mathbf{y} | \mathcal{E})$  is

$$\mathbf{x}, \mathbf{y} | \mathcal{E} \sim \mathcal{N}([\alpha_{\mathbf{x}}, \alpha_{\mathbf{y}}], \Sigma), \tag{1}$$

where

$$\Sigma = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix}. \tag{2}$$

The notation is statistical: Eq. (1) denotes that  $(\mathbf{x}, \mathbf{y})$  are random and normally distributed ( $\mathcal{N}$ ) with a mean  $[\alpha_{\mathbf{x}}, \alpha_{\mathbf{y}}]$  and covariance  $\Sigma$  as conditioned by ensemble  $\mathcal{E}$ . The posterior probability density function  $\text{Po}(\mathbf{y} | \mathbf{d}, \mathcal{E})$  after modification by data  $\mathbf{d}$  is obtained by application of Bayes's theorem:

$$\text{Po}(\mathbf{y} | \mathbf{d}, \mathcal{E}) \propto \int d\mathbf{x} P_L(\mathbf{d} | \hat{\mathbf{x}} = \mathbf{x}) P(\mathbf{x}, \mathbf{y} | \mathcal{E}), \tag{3}$$

in which  $P_L(\mathbf{d} | \hat{\mathbf{x}})$  is the Bayesian likelihood function. The likelihood function returns the probability that the data  $\mathbf{d}$  would have been obtained if the actual climate realized the state  $\hat{\mathbf{x}}$  as might be produced by a climate model drawn from a population described by Eq. (1).

The formulation of the likelihood function depends on the specific definition of the hindcast variable. If each member model of  $\mathcal{E}$  were to contribute its mean hindcast to the ensemble of hindcasts, then the data  $\mathbf{d}$  would be expected to differ from  $\hat{\mathbf{x}}$  by measurement error and internal variability, the latter requiring estimation by an initial condition ensemble of each model. Because few of the members of the CMIP5 ensemble offer initial condition ensembles and those that do contribute only small initial condition ensembles, this definition of  $\mathbf{x}$  is ill suited to our purposes. If, on the other hand, each member of  $\mathcal{E}$  were to contribute only one hindcast run, the intermodel covariance of the hindcasts would reflect uncertainty resulting from physical differences and internal variability,  $\mathbf{x}$  would be defined as the actual evolution of the climate system which the data  $\mathbf{d}$  should match perfectly in the absence of measurement error. The ensemble would span uncertainty in model physics and model initial conditions, and either or both could be inferred given appropriate data. On short time scales for prediction, the likelihood function is more sensitive to initial conditions, and so model physics can be treated as a Bayesian “nuisance parameter.” (This is a qualitative description of modern numerical weather prediction with randomized physics.) On long time scales of prediction, the likelihood function is more sensitive to model physics, and so the initial conditions can be treated as nuisance parameters. The latter is the case

given the subject matter of this research, and so hereinafter  $\hat{\mathbf{x}}$  is considered a single hindcast realization for a climate model (and  $\mathbf{d}$  a single realization of reality). As required of Bayesian nuisance parameters, we integrate over initial conditions by considering a variety of initial conditions within reasonable uncertainty as is explicitly done by the different member models of CMIP5. Ultimately, when testing an individual climate model with actual data, it is necessary to estimate a mean hindcast of the climate model and consider both internal variability and measurement error as the width of the likelihood function. Here we ignore measurement error in prioritizing tests of climate models. The addition of measurement error is unlikely to upset the relative priority of different tests of climate models inasmuch as measurement error should be a fraction of internal variability for climate monitoring missions (Leroy et al. 2008b).

The posterior distribution  $\text{Po}(\mathbf{y} | \mathbf{d}, \mathcal{E})$  is Gaussian:

$$\mathbf{y} | \mathbf{d}, \mathcal{E} \sim \mathcal{N}(\beta_{\mathbf{y}}, \Pi_{\mathbf{yy}}) \tag{4}$$

in which the most probable prediction is  $\beta_{\mathbf{y}}$  and its uncertainty covariance is  $\Pi_{\mathbf{yy}}$ . They are given by

$$\beta_{\mathbf{y}} = \alpha_{\mathbf{y}} + \Sigma_{\mathbf{yx}} \Sigma_{\mathbf{xx}}^{-1} (\mathbf{d} - \alpha_{\mathbf{x}}) \quad \text{and} \tag{5a}$$

$$\Pi_{\mathbf{yy}} = \Sigma_{\mathbf{yy}} - \Sigma_{\mathbf{yx}} \Sigma_{\mathbf{xx}}^{-1} \Sigma_{\mathbf{xy}}. \tag{5b}$$

We consider the prediction variable  $\mathbf{y}$  to be a one-element vector, and so the value of the proposed observable  $\mathbf{x}$  for predicting climate change is proportional to  $|\Sigma_{\mathbf{yy}}|/|\Pi_{\mathbf{yy}}|$ . Equation (5b) reduces to

$$|\Pi_{\mathbf{yy}}| = |\Sigma_{\mathbf{yy}}| (1 - \rho^2) \tag{6}$$

if the prediction variable  $\mathbf{y}$  is a scalar  $y$ , in which  $\rho$  is the correlation coefficient between an element of the hindcast observable  $\mathbf{x}$  and climate prediction  $y$  across the ensemble  $\mathcal{E}$ . The greater the correlation coefficient  $\rho$ , the higher the priority of the hindcast observation type and the test metric.

We anticipate that climate models will always be improved in the context of known trends in surface air temperature over the same time period that a newer highly accurate satellite-based data type is considered. Climate modeling groups already test their models and tune them such that their hindcast decadal temperature trends agree with observed temperature trends over the same period. In this sense, the ensemble of climate models contributed to model intercomparison projects is “calibrated.” Different climate modeling groups, however, tune different model physics to bring their models into agreement with temperature trends, and these different tuning practices beget

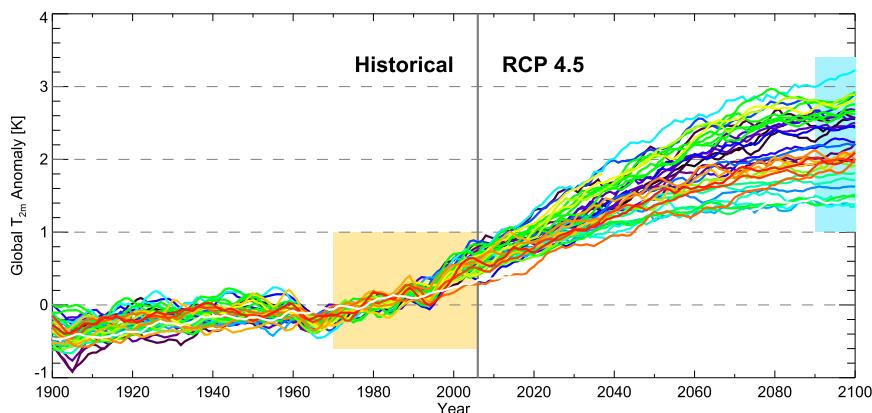


FIG. 1. The experimental design. This plot shows the evolution of global average surface air temperature from 1900 through 2100 generated by 40 CMIP5 models with 5-yr smoothing and with the temporal average of years 1950 through 2000 subtracted for each model. Temperature prior to 2006 is taken from a historical run of each model, and temperature after 2006 is taken from the corresponding RCP4.5 run of each model. Only the CMIP5 ensemble r1i1p1 was taken for each model. The observation types used in this study are simulated based on output of the historical runs from 1970 through 2005 (beige box). The climate predictions are changes in global average surface air temperature for the decade 2090–99 (blue box) from a best estimate for year 2006.

the remaining broad range of multidecadal climate predictions among the models that we are seeking to reduce. It is safe to assume that climate modeling groups will continue to calibrate their models according to trends in surface air temperature for years to come.

Figure 1 illustrates the experimental design. The hindcasts ( $\mathbf{x}$ ) are taken from the “historical” scenario of CMIP5, characterized by prescribed anthropogenic and natural radiative forcing from 1850 through 2005. For the hindcasts we produce temporal averages and trends over the period 1970 through 2005, a period in which models tend to produce a strong linear trend in global average surface air temperature. The hindcasts are resampled onto a  $2^\circ \times 2^\circ$  longitude–latitude grid such that regional averages are preserved. The climate predictions ( $\mathbf{y}$ ) are based on the output of the representative concentration pathway 4.5 (RCP4.5) scenario, characterized by  $4.5 \text{ W m}^{-2}$  of stabilized radiative forcing after 2100 over preindustrial conditions (Moss et al. 2010). It is considered a moderate projection of future radiative forcing. Multidecadal climate prediction is the difference between the 2090–99 average and global average surface air temperature and a best estimate for the global average surface air temperature in year 2006. The best estimate in 2006 is the intercept at year 2006 determined by linear regression of global average surface air temperature over the period 2006 through 2045. For each model we take only one run, which is connected seamlessly from its historical to its future RCP4.5 scenario, thus guaranteeing continuous time series.

Multidecadal climate prediction, by virtue of its time scale, is expected to be governed by transience in the ocean and climate sensitivity as dictated by atmospheric and surface feedbacks. Figure 2 shows the relationship between climate sensitivity and predicted multidecadal temperature change as defined above, as if climate sensitivity were an observation type ( $x$ ). The correlation between sensitivity and prediction is 0.56, pointing toward climate sensitivity as a controlling factor in multidecadal temperature prediction. The normality of the distribution in Fig. 2 is typical of all the distributions considered in this work.

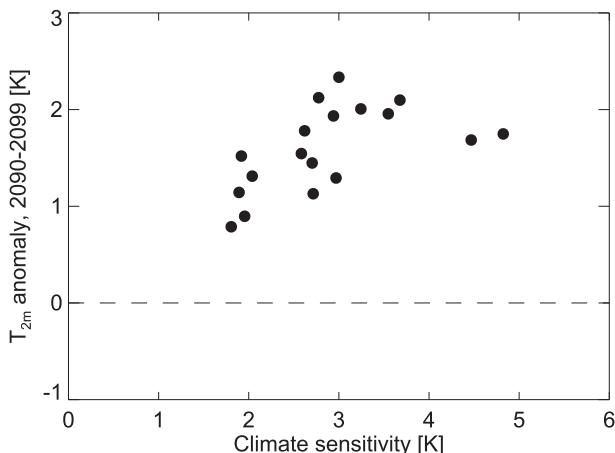


FIG. 2. Scatterplot of climate sensitivity vs multidecadal change in global average surface air temperature. Climate sensitivity was calculated as described in Andrews et al. (2012). The models are taken from the CMIP5 ensemble.

### 3. Remote sensing observable types

Three remote sensing observation types have either been deployed or technically developed for the purpose of climate monitoring, which means on-board traceability to international standards and adequate spatial–temporal coverage. They are radio occultation using the Global Navigation Satellite Systems (GNSS), high-spectral-resolution thermal infrared spectra (Anderson et al. 2004), and high-spectral-resolution reflected shortwave radiance (Wielicki et al. 2013). GNSS radio occultation has already been deployed in the form of missions of opportunity (Wickert et al. 2001) and as the Constellation Observing System for the Meteorology, Ionosphere, and Climate (COSMIC) constellation (Liou et al. 2007); high-accuracy, high-spectral-resolution longwave and shortwave instruments have been developed to the point of space-readiness but have not yet been deployed. For GNSS radio occultation we consider long-term averages and trends of the geopotential height of the 200-hPa surface, which can be retrieved directly from GNSS radio occultation data (Leroy 1997) and is strongly related to upper-air winds (Leroy et al. 2006; Verkhoglyadova et al. 2014). For the radiation fields we consider broadband downward shortwave and longwave radiative flux at the top of the atmosphere because net downward flux warms the climate system. The radiation fields are for all-sky conditions. Net downward longwave radiation is the opposite of outgoing longwave radiation (CMIP’s variable rlut), and net downward shortwave radiation is the difference between insolation (rsdt) and reflected shortwave flux at the top of the atmosphere (rsut). For diagnostic purposes we also consider temporal averages and trends in net total downward radiation at the top of the atmosphere (TOA) as hindcasts even though net total TOA radiation is extremely difficult to measure with desirable accuracy. The ability of long-term temporal averages and trends of these observable types to inform multidecadal climate prediction is examined in this section. Figures 3 and 4 show the covariances of hindcast and climate prediction ( $\Sigma_{xy}$ ) for temporal averages and trends of the prospective observation types and the spread in hindcast observations across the ensemble  $\mathcal{E}(\Sigma_{xx})$ , and Fig. 5 shows the squared correlation of hindcast and climate prediction ( $\rho^2$ ) for the same observation types.

The covariances  $\Sigma_{xy}$  shown in Fig. 3 offer easy physical insight. In those plots several large, coherent structures stand out. When the 200-hPa height field is the hindcast, structures related to position of the mid-latitude jet streams and warming of Arctic upper air are prominent; when the temporal average of the radiation fields are considered as hindcasts, structures related to

clouds and convection stand out; and when trends of the longwave and shortwave fields are considered, the dipole pattern in the equatorial Pacific, which can be explained as a trend in the strength of the Walker circulation, is prominent. Spatially coherent structures such as these in the covariance plots are expected. The hindcasts of climate models differ in ways that should be explainable by acknowledged uncertainties in model physics. Differences in physics emerge over broad regions that are influenced by those physics. If the patterns associated with the uncertain physics correlate strongly with climate predictions, then they may become important metrics for testing climate models.

The spread in hindcast observations  $\Sigma_{xx}$  across the ensemble are shown in Fig. 4. Because we retained only one model run for each member model of the ensemble (as discussed in the previous section), the spread in hindcast observations is explained by both intermodel differences in physics and internal variability. In the radiation fields, the spread in hindcast observations is dominated by clouds. For the average net TOA radiation, large intermodel differences are apparent in the equatorial Pacific and over regions of widespread stratocumulus clouds. For the average TOA shortwave, the intermodel differences of tropical and jet-stream clouds are apparent as well as intermodel differences of snow and ice. For the average TOA longwave, differences in clouds in the western and central Pacific are readily apparent. For trends in the radiation fields, both the shortwave and longwave show large differences in the western and central equatorial Pacific clouds, likely associated with trends in the Walker circulation. Differences in trends in shortwave and longwave roughly cancel when the two are added together to form the net TOA total radiation; hence, differences in trends in net TOA radiation are small.

The squared correlations shown in Fig. 5 suggest priority for hindcast observation types and their metrics when testing climate models for the multidecadal predictive ability. The squared correlation coefficient  $\rho^2$  given by

$$\rho^2 = \frac{|\Sigma_{xy}|^2}{\text{Diag}(\Sigma_{xx})|\Sigma_{yy}|} \quad (7)$$

constitute maps wherein the elements of  $\Sigma_{xy}$  correspond to individual longitude–latitude grid points, each divided by the diagonal element of  $\Sigma_{xx}$ , denoted  $\text{Diag}(\Sigma_{xx})$ , corresponding to the appropriate longitude–latitude grid point. The squared correlation is related to reduction in predictive uncertainty according to Eq. (6). Many of the plots in Fig. 5 exhibit squared correlations that exceed 0.50, implying the potential to reduce

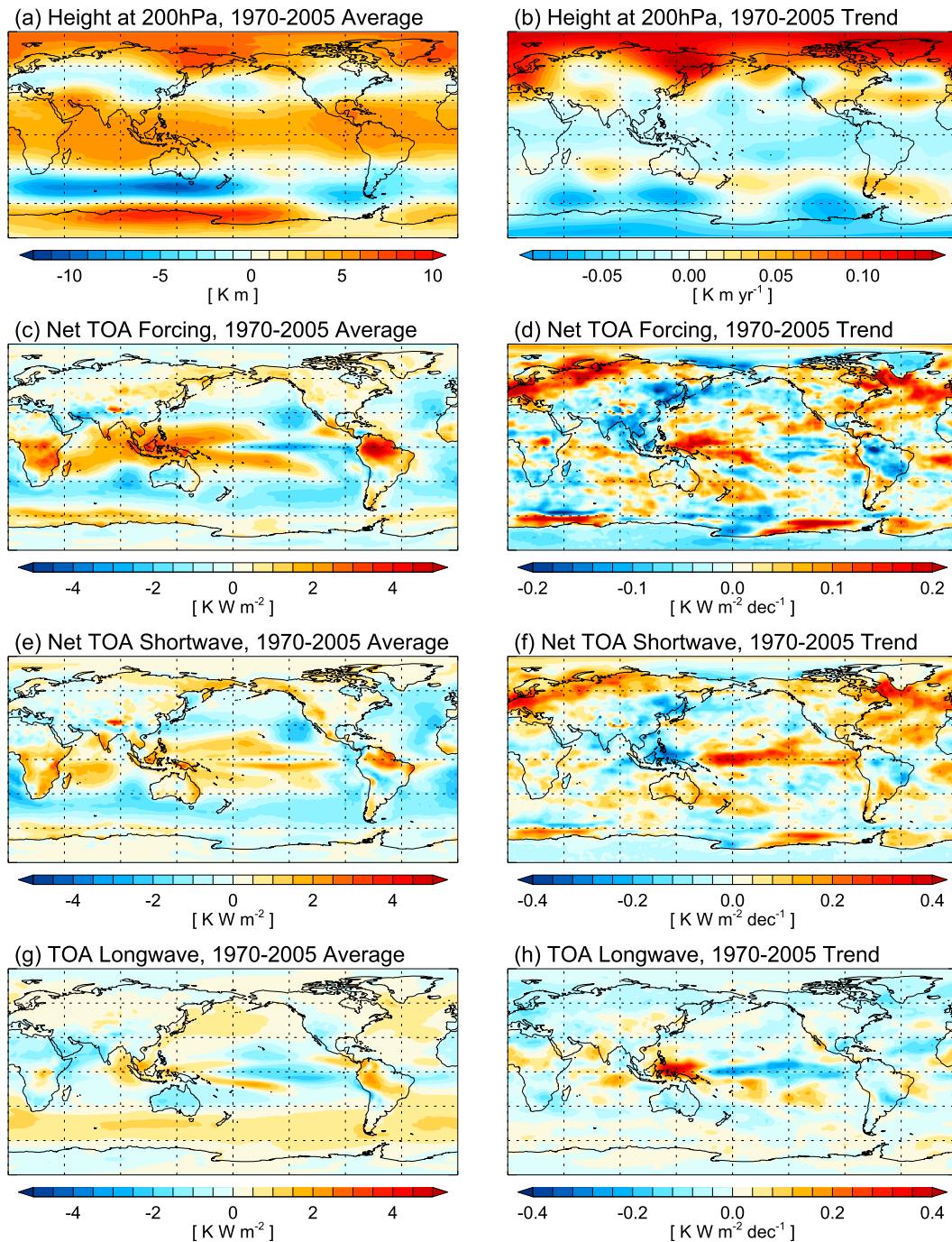


FIG. 3. The covariance of hindcast and climate prediction ( $\Sigma_{xy}$ ) considering several different observation types ( $\mathbf{x}$ ). The hindcasts are (a),(b) geopotential height of the 200-hPa surface, which is a proxy for GNSS radio occultation, (c),(d) net TOA downward radiation, (e),(f) net TOA downward shortwave radiation, and (g),(h) net TOA downward longwave radiation. Both (left) temporal averages over 1970–2005 and (right) trends over the same period are considered as hindcasts ( $\mathbf{x}$ ). The climate prediction ( $\mathbf{y}$ ) in each case is the change in global average temperature for the decade 2090–99 from a best estimate for the year 2006.

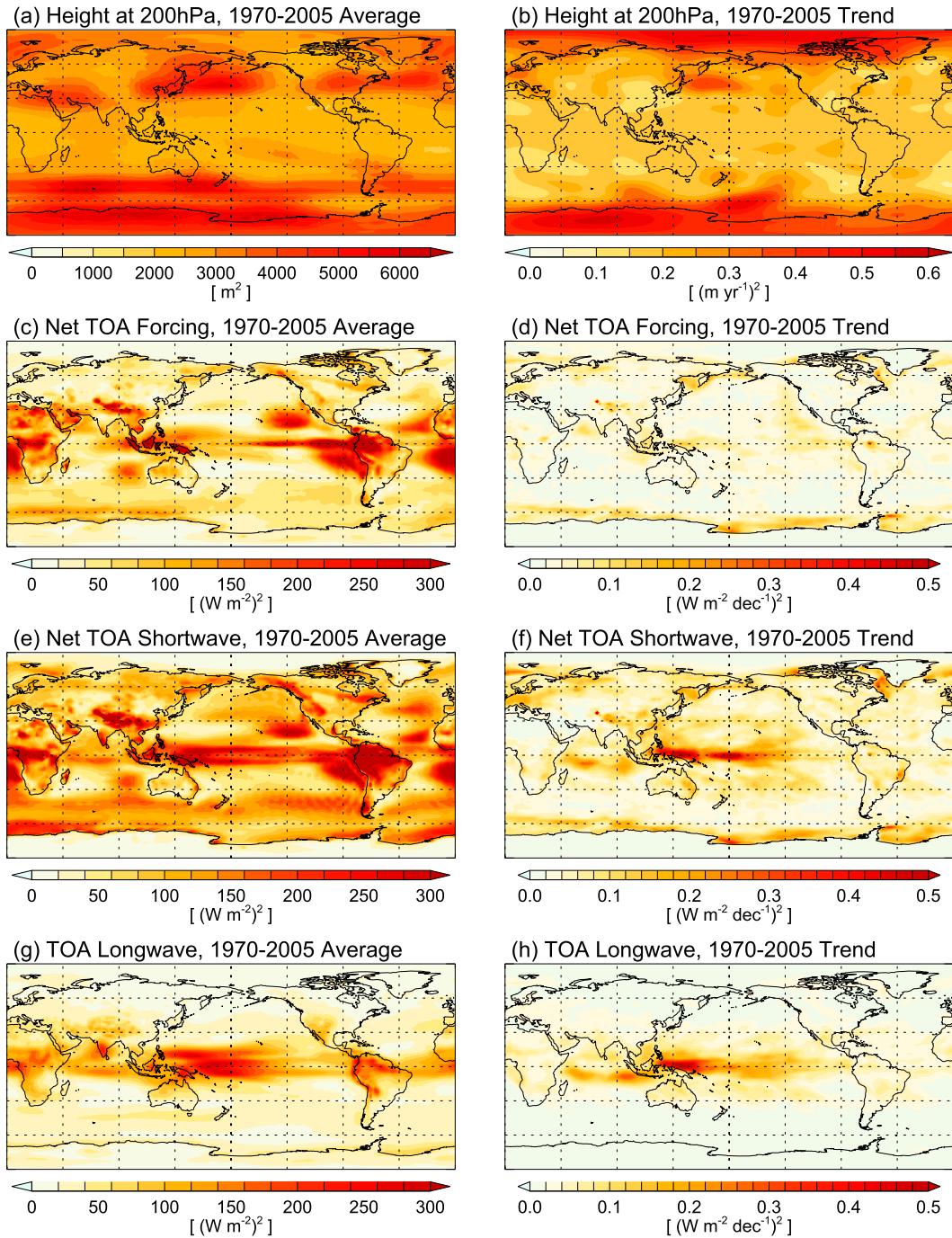


FIG. 4. The main diagonal of the covariance of the hindcast observations ( $\Sigma_{xx}$ ) considering several different observation types ( $\mathbf{x}$ ). The observation types are the same as in Fig. 3. The variances shown here are the intermodel differences in hindcast observations, including both differences in model physics and internal variability.

uncertainty in climate prediction by 30% (i.e.,  $1 - \sqrt{1 - \rho^2}$ ) given the right observation type and metric. On the other hand, even if no correlation actually exists between observation and climate prediction, “estimators” of the correlation are expected to be nonzero, and some large, because the ensemble of climate models used

to compute  $\Sigma_{xy}$  and  $\Sigma_{xx}$  is limited in size. The quantities contoured in Fig. 5 are estimators  $r$  of correlations  $\rho$ , with  $\rho$  understood as the correlation that would result should an infinite ensemble of climate models have been considered and  $r$  as the correlation that results from the limited draw  $\mathcal{E}$  (CMIP5) from the infinite ensemble.

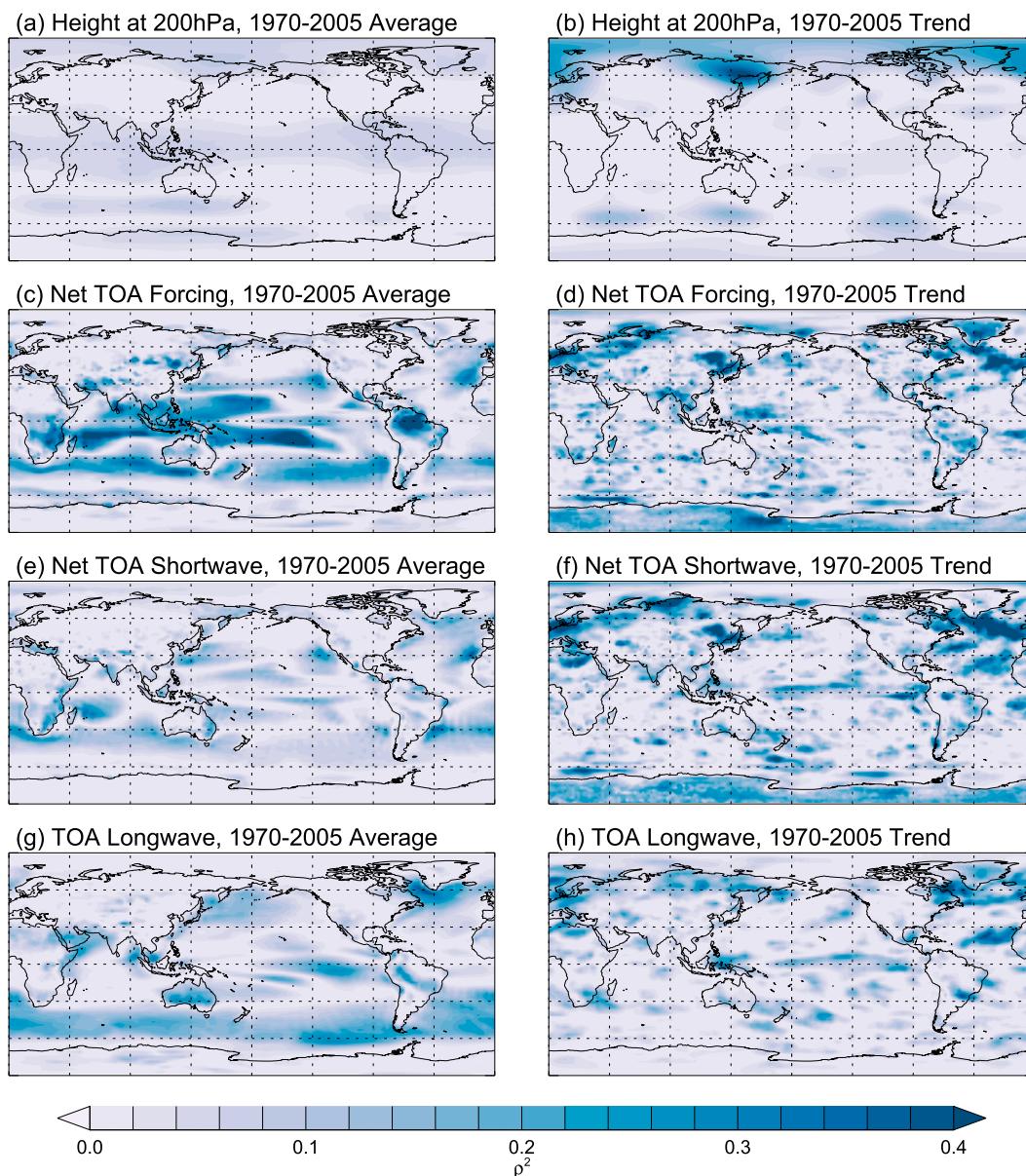


FIG. 5. The squared correlation of hindcast and climate prediction ( $\rho^2$ ) considering several different observation types. The hindcasts ( $\mathbf{x}$ ) and climate prediction ( $\mathbf{y}$ ) are the same as in Figs. 3 and 4.

A significance test must be applied to determine whether there are possibly nonzero correlations present.

If  $\rho = 0$  everywhere on the globe, the expected probability density function for estimators of correlation  $r$  follow a Student's  $t$  distribution with

$$t = r \left( \frac{1 - r^2}{n_{\text{eff}} - 2} \right)^{-1/2}, \quad (8)$$

where  $n_{\text{eff}}$  is degrees of freedom in the computation of  $r$  (Alder and Roessler 1972). In this case, it also corresponds

to the number of “effectively independent” models used to compute the correlation coefficients in each map. While approximately 50 models have contributed to CMIP5, not all have contributed all of the output variables to the historical and RCP4.5 scenarios needed for hindcast and prediction variables in this work. Consequently, we have typically used only 25 of the CMIP5 models to compute  $r$ . Many of the entries in the CMIP5 ensemble have physics in common, however, and so the number of effectively independent models is less than the number of models used to compute the correlation coefficients. Figure 6 shows plots of histograms of the

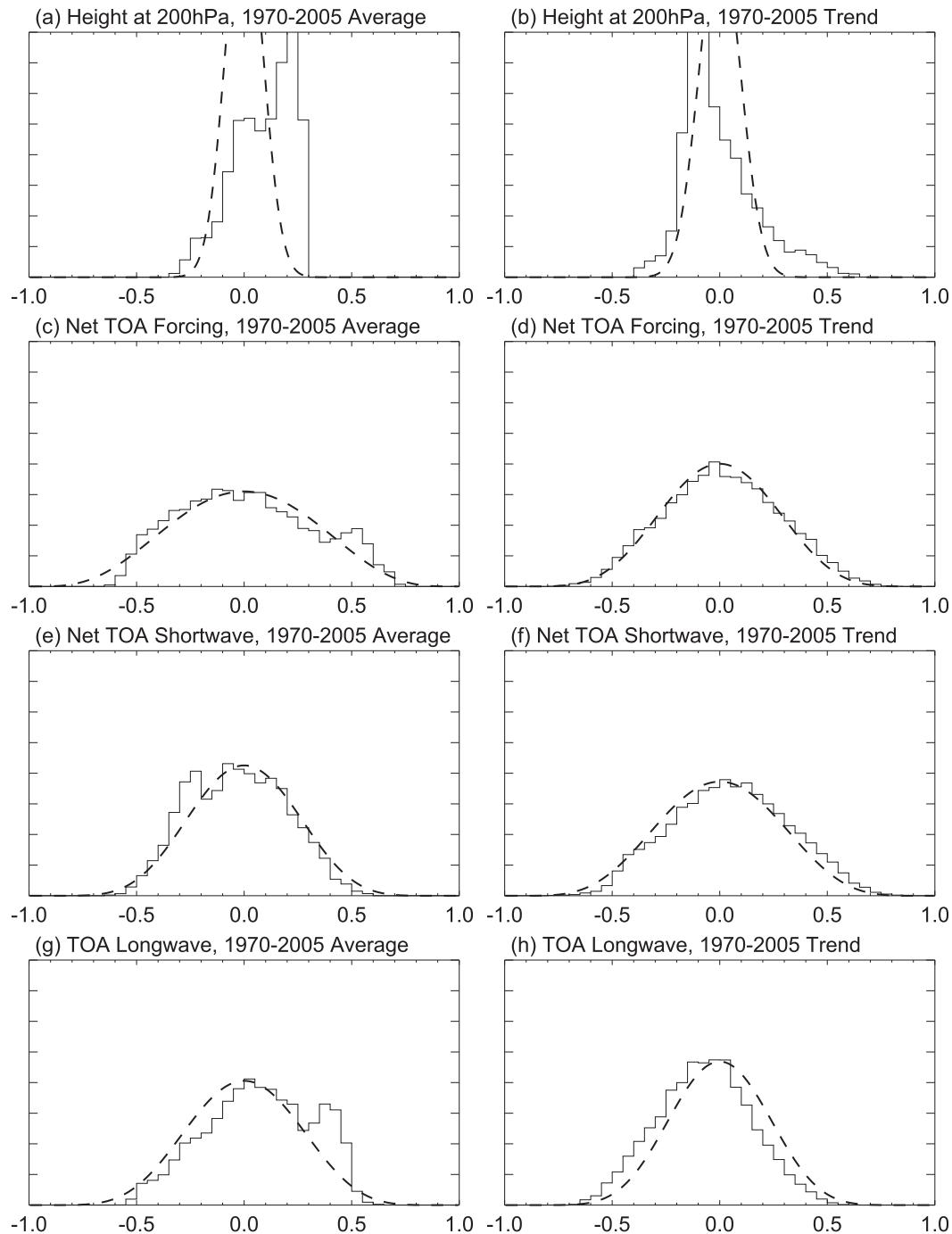


FIG. 6. Histograms of the estimators of area-weighted correlation coefficient ( $r$ ) across the contour plots of Fig. 5. The hindcasts are the same as in Fig. 3. The dashed line is the distribution expected of  $r$  if there were zero correlation ( $\rho = 0$ ) everywhere on the globe but with the effective number of degrees of freedom ( $n_{\text{eff}}$ ) chosen to fit the maximum of each histogram except where  $n_{\text{eff}}$  exceeds the number of models used for determine the correlation coefficients.

estimators of correlation coefficients taken from every longitude–latitude grid point of the maps of Fig. 5, each point contributing its area on the sphere to the histogram. For each hindcast observation type, the single

parameter of the Student's  $t$  distribution  $n_{\text{eff}}$  that best fits the structure of each histogram is inferred by matching the maximum of the Student's  $t$  distribution and the maximum of the histogram of  $r$ . In this way we

circumvent difficulties that arise from the large coherent structures of Fig. 5 that appear as bumps superimposed on a background distribution. The number of effectively independent models  $n_{\text{eff}}$  is about 17 for most of the histograms of Fig. 6.

The number of data points that enter into the histograms of Fig. 6 is the number of grid elements in the maps of Fig. 5, but the effective number of data points is less because of spatial coherence in the plots of Fig. 5. The greater the spatial heterogeneity in a plot of Fig. 5 is, the more closely the corresponding histogram in Fig. 6 should match the Student's  $t$  distribution for zero correlation if no correlation were present. The larger the coherent structures in a plot of Fig. 5 are, the fewer the number of independent evaluations of  $r$ , resulting in a coarser correspondence to the Student's  $t$  distribution in Fig. 6 if no correlation were present.

A comparison of the Student's  $t$  distributions for zero correlation with the histograms of estimators of the correlation coefficients indicates general consistency with no nonzero correlation being present. Each of the global-scale coherent structures prominent in the plots of Fig. 5 contributes one of the aforementioned bumps in Fig. 6. They are particularly noticeable in Figs. 6c, 6e, and 6g, the time-average TOA radiation fields as hindcast observation types. Whether these are statistically significant correlations requires a more sophisticated analysis, which is presented in the following section.

#### 4. Eigenmode decomposition of hindcast data

Metrics for testing climate models using climate monitoring observation types almost certainly point toward eigenmode decomposition rather than point-by-point comparison, because the “modes” of differences between climate models' hindcasts very likely correspond to differences in the models' physics. One can expect the signature of differing model physics to occur on global scales. The eigenmodes  $\mathbf{e}_\mu$  of intermodel differences in hindcast over the ensemble  $\mathcal{E}$  are the area-weighted eigenvectors of the hindcast covariance matrix  $\Sigma_{\mathbf{xx}}$ . The eigenmodes are orthogonal, and so intuiting the physics responsible based on the spatial patterns of an individual eigenmode is impossible (Monahan et al. 2009). Nonetheless,  $m$  statistically significant eigenmodes should mean that  $m$  uncertainties in model physics are being explained. Here we examine whether or not those explained differences in model physics also play a role in intermodel differences in climate prediction.

The eigenmodes of the hindcast data type are best interpreted as the dominant modes that explain the

intermodel differences in simulations of the hindcast data:

$$\langle \mathbf{e}_\mu, \Sigma_{\mathbf{xx}} \rangle = \lambda_\mu \mathbf{e}_\mu \quad \text{and} \quad (9a)$$

$$\langle \mathbf{e}_\mu, \mathbf{e}_\nu \rangle = \delta_{\mu\nu}, \quad (9b)$$

where the inner product rule  $\langle \cdot, \cdot \rangle$  is an area-weighted global integral of the product of the arguments. The first four eigenmodes for hindcast trend in the 200-hPa height field, in the hindcast trend in net TOA shortwave radiation, and in the hindcast long-term average in net TOA longwave radiation are shown in Fig. 7. Here, the maps show the eigenmode multiplied by the square root of the eigenvalue  $\lambda_\mu$ , and so the contours have the same dimensions as the hindcast observable  $\mathbf{x}$ .

The correlation coefficient between eigenmode  $\mu$  of the hindcast observation and climate prediction is a scalar:

$$r_\mu = \langle \mathbf{e}_\mu, \Sigma_{\mathbf{xy}} \rangle (\lambda_\mu |\Sigma_{\mathbf{yy}}|)^{-1/2}. \quad (10)$$

The square of the modal correlation coefficient  $r_\mu$  is the fraction of the variance of the climate predictions  $y$  in ensemble  $\mathcal{E}$  explained by eigenmode  $\mu$ , and a sum of the squares of modal correlation coefficients gives a cumulative squared correlation that can be inserted into Eq. (6) to estimate reduction of uncertainty in climate prediction. With  $n_m$  independent models, however, all of the variance in climate model prediction is explained by  $n_m - 1$  eigenmodes,

$$\sum_{\mu=1}^{n_m-1} r_\mu^2 = 1, \quad (11)$$

which seems to suggest reduction of uncertainty in climate prediction to zero, an absurd result. Just as  $r$  is an estimator of  $\rho$ , however,  $r_\mu$  is an estimator for  $\rho_\mu$ , as if the  $n_m$  models that define ensemble  $\mathcal{E}$  were drawn from an infinite pool of models. The difficulty arises because the number of models  $n_m$  used to construct the covariance of hindcast uncertainty  $\Sigma_{\mathbf{xx}}$  is less than the rank of  $\Sigma_{\mathbf{xx}}$ .

A truncation criterion is needed to limit the number of eigenmodes that can be legitimately retained in uncertainty reduction. There are two possible approaches to limiting the number of eigenmodes: consideration of significance in associated variance  $\lambda_\mu$  and consideration of uncertainty in the associated modal correlation coefficient  $r_\mu$ . The criterion for significance in associated variance was presented by North et al. (1982). This criterion holds that the only statistically significant eigenmodes are those with eigenvalues significantly

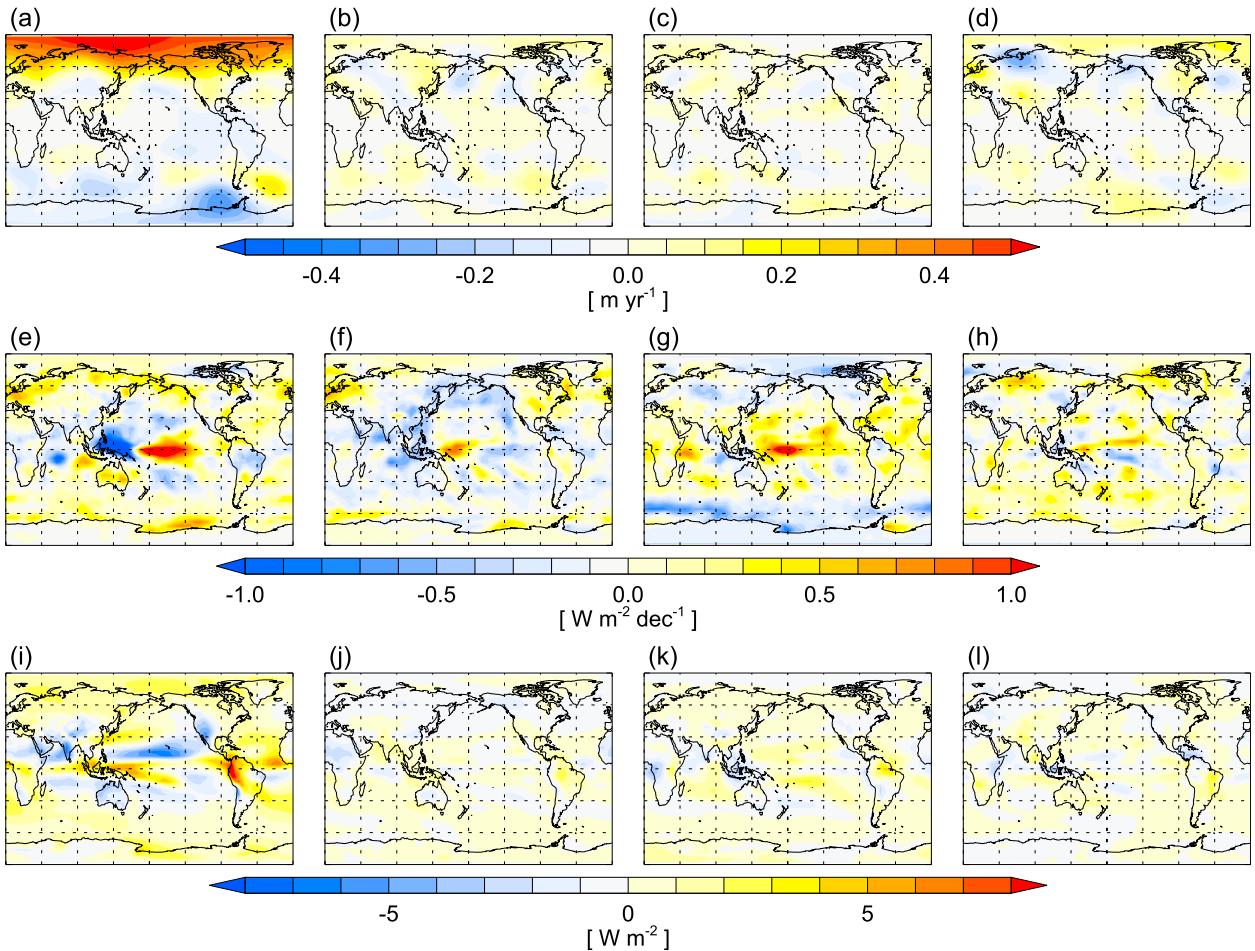


FIG. 7. Four eigenmodes for (a)–(d) the trend in the 200-hPa height as hindcast, (e)–(h) the trend in net TOA downward shortwave as hindcast, and (i)–(l) for the average in net TOA downward longwave as hindcast, sorted according to descending modal correlation coefficient  $|r_\mu|$ . The plots actually contain the eigenmodes multiplied by the square roots of their respective eigenvalues  $\lambda_\mu^{1/2} \mathbf{e}_\mu$ .

different from zero. The uncertainty in any eigenvalue is the sum of all of the eigenvalues divided by  $(n_m - 2)^{1/2}$ . Therefore, all of those eigenmodes whose eigenvalues explain more than  $(n_m - 2)^{-1/2}$  of the total of the eigenvalues of  $\Sigma_{\mathbf{xx}}$  can be considered significant modes of intermodel difference in hindcasts. Table 1 contains a list of the eigenvalues  $\lambda_\mu$  and modal correlation coefficient  $r_\mu$  for the gravest four modes of intermodel hindcast difference, sorted by decreasing eigenvalue  $\lambda_\mu$ . Application of the criterion of North et al. (1982) yields the conclusion that only one mode of the six hindcast observation types considered is statistically significant.

Significance of the modal correlation coefficient  $r_\mu$  is also considered. Table 2 contains a list of estimates of modal correlation coefficients for the eigenmodes of hindcast differences for the four eigenmodes sorted by descending  $|r_\mu|$ . We consider the  $r_\mu$  as estimators of the modal correlation coefficients  $\rho_\mu$  of an infinite pool of climate models but based on a subset of  $n_m$  draws taken

from the infinite pool. Because even  $n_m$  estimators  $r_\mu$  must explain all of the variance in the climate predictions of ensemble  $\mathcal{E}$  according to Eq. (11), it is possible that some of the  $|r_\mu|$  are very large even if all  $\rho_\mu$  of the infinite pool are zero. Figure 8 shows probability density functions for the largest  $r_\mu$  that can be expected for a given data type [ $\max(r_\mu^2)$ ] should  $n_m$  independent models have been used in the formulation of  $\Sigma_{\mathbf{xx}}$  and all  $\rho_\mu = 0$ . This was done by constructing 500 000 scenarios for each value of  $n_m$ . Each member of the 500 000 scenarios contains model principal components  $p_{\mu,i}$  and climate predictions  $y_i$ . The principal component  $p_{\mu,i}$  is best interpreted as the projection of model  $i$ 's hindcast observation field  $\mathbf{x}_i$  onto eigenmode  $\mathbf{e}_\mu$ :

$$p_{\mu,i} \equiv \langle \mathbf{e}_\mu, (\mathbf{x}_i - \bar{\mathbf{x}}) \rangle, \quad (12)$$

where  $\bar{\mathbf{x}} = (1/n_m) \sum_{i=1}^{n_m} \mathbf{x}_i$ . Values of  $p_{\mu,i}$  are simulated by first generating an  $(n_m + 1)$ -element vector  $\mathbf{x}_i$  for each

TABLE 1. Modal variance and modal correlation coefficient ( $|\rho_{\mu}|$ ) associated with the first four eigenmodes for various hindcast observation types, sorted by decreasing modal variance. Each hindcast observation type is either a temporal average or a trend over the years 1970–2005, as noted. The modal variance  $\lambda_{\mu}$  is given as a fraction of the total variance  $\sum_{\mu} \lambda_{\mu}$ . The modes are the eigenmodes of  $\Sigma_{xx}$ . The correlation coefficient  $r_{\mu}$  is defined according to Eq. (10).

	Mode 1	Mode 2	Mode 3	Mode 4
Hindcast type	$\lambda_{\mu}, r_{\mu}$	$\lambda_{\mu}, r_{\mu}$	$\lambda_{\mu}, r_{\mu}$	$\lambda_{\mu}, r_{\mu}$
Height (200 hPa), 1970–2005 average	0.64, 0.095	0.16, 0.087	0.09, 0.406	0.06, 0.254
Height (200 hPa), 1970–2005 trend	0.50, 0.017	0.12, 0.038	0.08, 0.582	0.06, 0.216
Shortwave, 1970–2005 average	0.33, 0.285	0.15, 0.235	0.12, 0.023	0.10, 0.205
Shortwave, 1970–2005 trend	0.17, 0.386	0.11, 0.371	0.09, 0.380	0.07, 0.308
Longwave, 1970–2005 average	0.30, 0.131	0.18, 0.276	0.13, 0.203	0.08, 0.499
Longwave, 1970–2005 trend	0.20, 0.312	0.11, 0.309	0.07, 0.119	0.06, 0.166

model  $i$  using a normal distribution random number generator and then determining the principal components  $p_{\mu,i}$  by singular value decomposition. As is standard for principal components,  $\sum_{i=1}^{n_m} p_{\mu,i} p_{\nu,i} = \lambda_{\mu} \delta_{\mu,\nu}$ .

The climate predictions  $y_i$  are also generated using a normal distribution random number generator, and the correlation coefficient estimators are computed by

$$r_{\mu} = \left[ \sum_{i=1}^{n_m} (p_{\mu,i} - \bar{p}_{\mu})^2 \sum_{i=1}^{n_m} (y_i - \bar{y})^2 \right]^{-1/2} \sum_{i=1}^{n_m} (p_{\mu,i} - \bar{p}_{\mu})(y_i - \bar{y}), \quad (13)$$

where  $\bar{p}_{\mu} = (1/n_m) \sum_{i=1}^{n_m} p_{\mu,i}$  and  $\bar{y} = (1/n_m) \sum_{i=1}^{n_m} y_i$ . The number of eigenmodes is  $n_m - 1$  for an ensemble of  $n_m$  models, thus  $\mu \in [1, n_m - 1]$ . For each scenario of the 500 000 scenarios, the largest value of  $r_{\mu}^2$  is retained and binned into the histograms of Fig. 8. Table 3 shows the values of  $\max(r_{\mu}^2)$  for which the hypothesis that all  $\rho_{\mu} = 0$  can be rejected with varying degrees of confidence. These values are obtained by integrating under the curves of Fig. 8. The relevant number of degrees of freedom when assessing statistical significance of correlation coefficients is the effective number of degrees of freedom, previously found to be approximately 17.

Only three eigenmodes in all the considered hindcast observation types have considerable fractional variance ( $\lambda_{\mu} > 0.01$ ) and correlate significantly with climate prediction ( $r_{\mu}$ ) even if marginally, as is apparent in Table 2. The three hindcast observation types are the 1970–2005 trend in the 200-hPa height, the 1970–2005 trend in

net TOA downward shortwave radiation, and the 1970–2005 average in net downward longwave radiation. The most prominent is the 1970–2005 trend in the 200-hPa height, with  $r_{\mu} = 0.582$ . The most relevant modes, as defined by large  $|r_{\mu}|$ , are plotted in Fig. 7. The confidence level that the eigenmode in Fig. 7a has a nonzero correlation with climate prediction is approximately 80%.

## 5. Discussion

We have pursued the prioritization of prospective remotely sensed climate monitoring observation types with the goal of improving climate prediction capability, having expanded upon a previous study by considering spatial information and long-term trends versus long-term averages over the same time series. We have performed calculations against a realistic backdrop, namely that climate models will continue to be calibrated with

TABLE 2. Modal correlation coefficient ( $|r_{\mu}|$ ) and modal variance ( $\lambda_{\mu}$ ) associated with the first four eigenmodes for various hindcast observation types, sorted by decreasing modal correlation coefficients (see Table 1). The eigenmodes have been renumbered after resorting.

	Mode 1	Mode 2	Mode 3	Mode 4
Hindcast type	$r_{\mu}, \lambda_{\mu}$	$r_{\mu}, \lambda_{\mu}$	$r_{\mu}, \lambda_{\mu}$	$r_{\mu}, \lambda_{\mu}$
Height (200 hPa), 1970–2005 average	0.520, 0.00	0.406, 0.09	0.343, 0.00	0.254, 0.06
Height (200 hPa), 1970–2005 trend	0.582, 0.08	0.374, 0.01	0.313, 0.00	0.282, 0.02
Shortwave, 1970–2005 average	0.590, 0.01	0.381, 0.01	0.339, 0.03	0.285, 0.04
Shortwave, 1970–2005 trend	0.386, 0.17	0.380, 0.09	0.371, 0.11	0.361, 0.06
Longwave, 1970–2005 average	0.499, 0.08	0.358, 0.01	0.321, 0.02	0.298, 0.01
Longwave, 1970–2005 trend	0.412, 0.05	0.351, 0.05	0.312, 0.20	0.309, 0.11

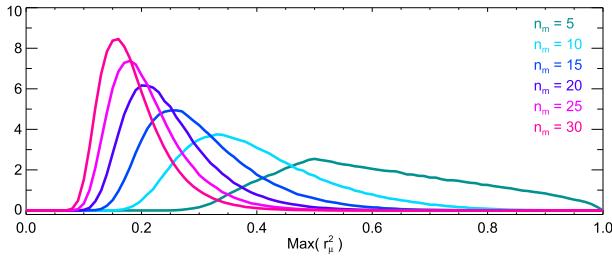


FIG. 8. The probability density functions for the maximum squared modal correlation coefficient ( $r_\mu$ ) for varying numbers of models ( $n_m$ ) drawn from an infinite pool with zero correlation between hindcast and climate prediction for all modes. The modal correlation coefficients are computed according to Eq. (11). Because the number of models  $n_m$  is less than the rank of the hindcast  $\mathbf{x}$ , the sum of all  $r_\mu^2$  is unity.

some existing data as they are currently. We anticipate that climate modeling groups will perform additional tests of their models against data, selecting first those data that they deem most important for testing their models. The models they generate will be tuned to agree with those data. In prioritizing observation types and metrics, we are recommending to climate modeling groups what data should be given priority in testing their models and what metrics should be applied in the tests. These observation types should be given priority in a climate monitoring system.

Using a method that has been applied to conditioning large perturbed physics ensembles of a climate model, we sought the hindcast observation types that correlate best with multidecadal change in global average surface air temperature. When resolved spatially, the hindcast observations can seemingly correlate very strongly with long-term climate prediction; however, when many thousands of correlations are estimated, statistical noise is capable of generating occasionally large estimates of correlation coefficients. We have shown that the correlations between hindcast and climate prediction are generally consistent with there being no nonzero correlation present (cf. Fig. 6) with the possible exception of a few prominent large-scale structures in Fig. 5. The candidate hindcast observation types for highest priority are Figs. 7a, 7e, and 7i. Figures 7a and 7e lend themselves to simple physical interpretation: tropospheric expansion in the Arctic as measured by GNSS radio occultation contains multidecadal predictive capability, and an increase in the strength of the Walker circulation as born out in net TOA downward shortwave radiation. The Walker circulation can increase low cloud fraction in the central and eastern tropical Pacific because of enhanced subsidence and decrease low cloud fraction in the tropical western Pacific by enhancing rising motion, leading to the signature seen in Fig. 7e. Change in the

TABLE 3. The values of  $\max(r_\mu^2)$  for which all  $\rho_\mu = 0$  can be rejected with confidences of 50%, 75%, 95%, and 99% for varying numbers of models  $n_m$  (see Fig. 8).

$n_m$	Confidence levels			
	50%	75%	95%	99%
5	0.597	0.736	0.906	0.968
10	0.383	0.474	0.639	0.754
15	0.292	0.361	0.493	0.599
20	0.240	0.294	0.403	0.495
25	0.205	0.251	0.344	0.425
30	0.180	0.220	0.301	0.373

intensity of the Walker circulation has attracted attention previously (Vecchi et al. 2006).

There are several possible causes for the slight improvements in climate prediction we have found. First, it is possible that the CMIP5 ensemble of climate models is not large enough to arrive at decisive statistics. Approximately 50 models contributed to CMIP5, about 25 of which contributed the scenarios needed in this study, and only about 17 of those can be considered independent. A large perturbed physics ensemble of a single climate model may provide better statistics, but previous efforts with perturbed physics ensembles have had difficulty reducing uncertainty in equilibrium climate sensitivity (Sexton et al. 2012). Second, the climate modeling groups may already have effectively tested their models against data corresponding to the observation types we have considered here. If so, then no correlation between hindcasts of that observation type and climate prediction would be found. This seems unlikely, though, because of the large spreads that exist in most of the hindcast observation types as expressed through  $\Sigma_{\mathbf{xx}}$ . Third, the spectral information contained in the radiance observation types being considered for the Climate Absolute Radiance and Refractivity Observatory (CLARREO) mission may enable the distinguishing between external radiative forcing and radiative response, and that separation might enable larger correlation coefficients. Radiative response should provide information on radiative feedbacks and hence equilibrium climate sensitivity (Leroy et al. 2008a; Huang et al. 2010; Feldman et al. 2013). On the other hand, the spatial structure of signals in hindcasts can be expected to contain much of the same information as spectral structure: there is little doubt in Fig. 3 where clouds are responsible for correlation and where ice is responsible. In any case, further investigation is warranted to understand multidecadal predictability of climate and to find climate monitoring observation types of the highest priority.

*Acknowledgments.* We acknowledge the World Climate Research Programme's Working Group on Coupled

Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. The authors thank Profs. Richard Goody and Guido Visconti for their intellectual contributions to this research. Dr. Leroy was supported by Grant NNX11AD01G of the U.S. National Aeronautics and Space Administration. Drs. Redaelli and Grassi were supported in part by the Italian Space Agency (ASI) under the QUILTSAT project.

## REFERENCES

- Alder, H., and E. Roessler, 1972: *Introduction to Probability and Statistics*. W. H. Freeman, 373 pp.
- Anderson, J., J. Dykema, R. Goody, H. Hu, and D. Kirk-Davidoff, 2004: Absolute, spectrally-resolved, thermal radiance: A benchmark for climate monitoring from space. *J. Quant. Spectrosc. Radiat. Transf.*, **85**, 367–383, doi:10.1016/S0022-4073(03)00232-2.
- Andrews, T., J. Gregory, M. Webb, and K. Taylor, 2012: Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. *Geophys. Res. Lett.*, **39**, L09712, doi:10.1029/2012GL051607.
- Feldman, D., C. Algieri, W. Collins, Y. Roberts, and P. Pilewskie, 2011: Simulation studies for the detection of changes in broadband albedo and shortwave nadir reflectance spectra under a climate change scenario. *J. Geophys. Res.*, **116**, D24103, doi:10.1029/2011JD016407.
- , D. Coleman, and W. Collins, 2013: On the usage of spectral and broadband satellite instrument measurements to differentiate climate models with different cloud feedback strengths. *J. Climate*, **26**, 6561–6574, doi:10.1175/JCLI-D-12-00378.1.
- Forest, C., P. Stone, A. Sokolov, M. Allen, and M. Webster, 2002: Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science*, **295**, 113–117, doi:10.1126/science.1064419.
- Geoffroy, O., D. Saint-Martin, and A. Ribes, 2012: Quantifying the sources of spread in climate change experiments. *Geophys. Res. Lett.*, **39**, L24703, doi:10.1029/2012GL054172.
- Hall, A., and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.*, **33**, L03502, doi:10.1029/2005GL025127.
- Huang, Y., S. Leroy, P. Gero, J. Dykema, and J. Anderson, 2010: Separation of longwave climate feedbacks from spectral observations. *J. Geophys. Res.*, **115**, D07104, doi:10.1029/2009JD012766.
- , —, and R. Goody, 2011: Discriminating between climate observations in terms of their ability to improve an ensemble of climate predictions. *Proc. Natl. Acad. Sci. USA*, **108**, 10 405–10 409, doi:10.1073/pnas.1107403108.
- Leroy, S., 1997: Measurement of geopotential heights by GPS radio occultation. *J. Geophys. Res.*, **102**, 6971–6986, doi:10.1029/96JD03083.
- , J. Anderson, and J. Dykema, 2006: Testing climate models using GPS radio occultation: A sensitivity analysis. *J. Geophys. Res.*, **111**, D17105, doi:10.1029/2005JD006145.
- , —, —, and R. Goody, 2008a: Testing climate models using thermal infrared spectra. *J. Climate*, **21**, 1863–1875, doi:10.1175/2007JCLI2061.1.
- , —, and G. Ohring, 2008b: Climate signal detection times and constraints on climate benchmark accuracy requirements. *J. Climate*, **21**, 841–846, doi:10.1175/2007JCLI1946.1.
- Liou, Y., A. Pavelyev, S. Liu, A. Pavelyev, N. Yen, C. Huang, and C. Fong, 2007: FORMOSAT-3/COSMIC GPS radio occultation mission: Preliminary results. *IEEE Trans. Geosci. Remote Sens.*, **45**, 3813–3826, doi:10.1109/TGRS.2007.903365.
- Min, S., and A. Hense, 2006a: A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models. *Geophys. Res. Lett.*, **33**, L08708, doi:10.1029/2006GL025779.
- , and —, 2006b: A Bayesian assessment of climate change using multimodel ensembles. Part I: Global mean surface temperature. *J. Climate*, **19**, 3237–3256, doi:10.1175/JCLI3784.1.
- Monahan, A., J. Fyfe, M. Ambaum, D. Stephenson, and G. North, 2009: Empirical orthogonal functions: The medium is the message. *J. Climate*, **22**, 6501–6514, doi:10.1175/2009JCLI3062.1.
- Moss, R., and Coauthors, 2010: The next generation of scenarios for climate change research and assessment. *Nature*, **463**, 747–756, doi:10.1038/nature08823.
- North, G., T. Bell, R. Cahalan, and F. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **110**, 699–706, doi:10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2.
- Sexton, D., J. Murphy, M. Collins, and M. Webb, 2012: Multivariate probabilistic projections using imperfect climate models. Part I: Outline of methodology. *Climate Dyn.*, **38**, 2513–2542, doi:10.1007/s00382-011-1208-9.
- Sivia, D., 2006: *Data Analysis: A Bayesian Tutorial*. Oxford University Press, 246 pp.
- Taylor, K., R. Stouffer, and G. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-00094.1.
- Vecchi, G., B. Soden, A. Wittenberg, I. Held, A. Leetma, and M. Harrison, 2006: Weakening of the tropical Pacific atmospheric circulation due to anthropogenic forcing. *Nature*, **441**, 73–76, doi:10.1038/nature04744.
- Verkhoglyadova, O., S. Leroy, and C. Ao, 2014: Estimation of winds from GPS radio occultations. *J. Atmos. Oceanic Technol.*, **31**, 2451–2461, doi:10.1175/JTECH-D-14-00061.1.
- Wickert, J., and Coauthors, 2001: Atmosphere sounding by GPS radio occultation: First results from CHAMP. *Geophys. Res. Lett.*, **28**, 3263–3266, doi:10.1029/2001GL013117.
- Wielicki, B., and Coauthors, 2013: Achieving climate change absolute accuracy in orbit. *Bull. Amer. Meteor. Soc.*, **94**, 1519–1539, doi:10.1175/BAMS-D-12-00149.1.