Spatial-Scale Dependence of Climate Model Performance in the CMIP3 Ensemble

DAVID MASSON AND RETO KNUTTI

Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland

(Manuscript received 12 November 2009, in final form 3 February 2011)

ABSTRACT

About 20 global climate models have been run for the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) to predict climate change due to anthropogenic activities. Evaluating these models is an important step to establish confidence in climate projections. Model evaluation, however, is often performed on a gridpoint basis despite the fact that models are known to often be unreliable at such small spatial scales. In this study, the annual mean values of surface air temperature and precipitation are analyzed. Using a spatial smoothing technique with a variable-scale parameter it is shown that the intermodel spread, as well as model errors from observations, is reduced as the characteristic smoothing scale increases. At the same time, the ability to reproduce small-scale features is reduced and the simulated patterns become fuzzy. Depending on the variable of interest, the location, and the way that data are aggregated, different optimal smoothing scales from the gridpoint size to about 2000 km are found to give good agreement with present-day observation yet retain most regional features of the climate signal. Higher model resolution surprisingly does not imply much better agreement with temperature observations, in particular with stronger smoothing, and resolving smaller scales therefore does not necessarily seem to improve the simulation of large-scale climate features. Similarities in mean temperature and precipitation fields for a pair of models in the ensemble persist locally for about a century into the future, providing some justification for subtracting control errors in the models. Large-scale to global errors, however, are not well preserved over time, consistent with a poor constraint of the present-day climate on the simulated global temperature and precipitation response.

1. Introduction

To assess future climate changes and the anthropogenic contribution to global warming, about 20 global climate models ran scenarios in a coordinated model intercomparison targeted for the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4). These models came from various institutions and differ in their structure. While there is agreement between these models concerning the anthropogenic contribution to global warming and some agreement on the projected future changes, the uncertainty quantification is still problematic and a consensus on performance metrics for models is lacking (Tebaldi and Knutti 2007; Knutti et al. 2010b). Some questions related to model evaluation have rarely been asked. For example, what is the typical spatial scale at which models can provide reliable results? Climate scientists have an intuitive feeling for it and use it when interpreting results. However, models are also often compared to observations or other models at the gridpoint scale, sometimes leading to the conclusion that models do not even agree on the sign of predicted future changes over large areas on the globe. This is particularly true for variables other than temperature, for example, precipitation (see Fig. 10.12 of Meehl et al. 2007b). Evaluating the models directly on the smallest spatial scale can be misleading because resolving a feature requires several grid points at least. Therefore, errors on small spatial scales can be large even if the models agree better on larger scales. Apart from this numerical aspect, natural variability is also an important source of model disagreement. Aggregating changes over larger regions reduces internal variability (Räisänen 2001) and leads to more consistent projections across models even for variables that are difficult to simulate—for example, precipitation where zonal averaging leads to a more consistent pattern across the models (Zhang et al. 2007).

While much of the community's effort goes into improving the models (shown for example by Reichler and Kim 2008), it is unclear at which scale the models can provide useful information and agreement with data, how that scale depends on the variable and projection

Corresponding author address: David Masson, ETH Zurich, Universitaetstr. 16, CH-8092 Zurich, Switzerland.
E-mail: david.masson@env.ethz.ch

DOI: 10.1175/2011JCLI3513.1

lead time, and whether higher model resolution leads to more useful information on a smaller spatial scale (Stainforth et al. 2007). Between the gridpoint scale where models are less reliable and the global scale that is of limited use for local projections, a whole range of spatial scales exists and can be explored. Climate projections were often aggregated regionally to reduce model uncertainty (e.g., Tebaldi et al. 2005) but the regions were chosen in a rather ad hoc way. In the present paper, we attempt to estimate optimal spatial smoothing scales for temperature and precipitation in a more formal way by minimizing a penalty function, which is a combination of the model's error to an observation-based dataset and a measure of spatial information that is lost through averaging. In simple words, the full local information is provided at every grid point without smoothing, but model errors and model spread may be large and confidence in local projections is therefore low. On very large scales, errors are smaller and models are more likely to agree, but the information for local impacts is lost and the projection is again rather useless. Somewhere in between is a regional to continental aggregation or smoothing (implicit for example in Christensen et al. 2007) where information is most likely to be useful and robust against model assumptions.

We first test the agreement of present-day simulated climate with observations at different spatial scales. Then we focus on how model resolution affects model errors for different areas and scales. Next, we consider the ratio of the climate change signal to the model disagreement and how smoothing the data affects when and where models agree on a predicted change. Finally, we study the persistence of model errors of the initial period 1960–79 (or equivalently the similarity of two models in the ensemble) through time and for different spatial scales in a perfect model approach. This is an important point since past agreement with observations is often used to support projections into the future (Stott and Kettleborough 2002; Giorgi and Mearns 2002, 2003; Tebaldi et al. 2005; Knutti 2008b,a), that is, it is assumed that a model that is close to observations in the past will be close to the real world in its simulated future response.

2. Method

a. Data

This study uses a subset of the data produced for the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project phase 3 (CMIP3) (Meehl et al. 2007a), a coordinated model intercomparison for the AR4 report of the IPCC. One ensemble member for each of the 24 atmosphere–ocean general circulation models (AOGCMs) simulations under the A1B emission scenario (Nakicenovic et al. 2000) is used.

The observation-based datasets are the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) (Uppala et al. 2005) for surface temperature and the Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP) (Xie and Arkin 1997) and the Global Precipitation Climatology Project (GPCP) (Adler et al. 2003) for precipitation. Each of these observation-based datasets obviously has its limitations and errors, and they partly rely on models as well, but still these are probably the best observation-based datasets available for our purpose. Because individual models and observations come at different spatial resolutions, the data is interpolated using bilinear interpolation to a common T42 grid (Gaussian grid associated with spectral truncation, 128 latitudinal by 64 longitudinal grid points). The original data contains monthly averaged fields from which climatological averages over periods of 20 yr (annual means) have been extracted. For future periods where no observations exist, the perfect model approach is used in some cases, that is, each model is treated as reference once and the results are averaged afterward.

b. Field smoothing

To study uncertainty of models on different spatial scales, a field-smoothing technique is used. Instead of evaluating models at the gridpoint scale, the fields are smoothed by weighted spatial averaging, whereby the weight $w_{i,j}$ of each point (i, j) decreases exponentially with the squared distance $d_{i,j}(k, l)$ from the original location (k, l):

$$w_{i,j}(k, l, \lambda) = e^{-d_{i,j}^2(k,l)/2\lambda^2},$$

with λ being the parameter representing the characteristic smoothing length scale of the Gaussian weighting. Therefore, the original data $V_{k,l}$ is replaced by its smoothed value $\overline{V}_{k,l}$ according to

$$\overline{V}_{k,l}(\lambda) = \frac{\sum_{i,j}^{I,J} V_{i,j} w_{i,j}(k, l, \lambda)}{\sum_{i,j}^{I,J} w_{i,j}(k, l, \lambda)},$$

with I, J being the total number of longitude and latitude grid points. To characterize how smooth or homogeneous the climate signal is at a certain location, we introduce the spatial variation $\sigma^s(\lambda)$, which is essentially a standard deviation of all grid points weighted by $w_{i,j}$ defined above:

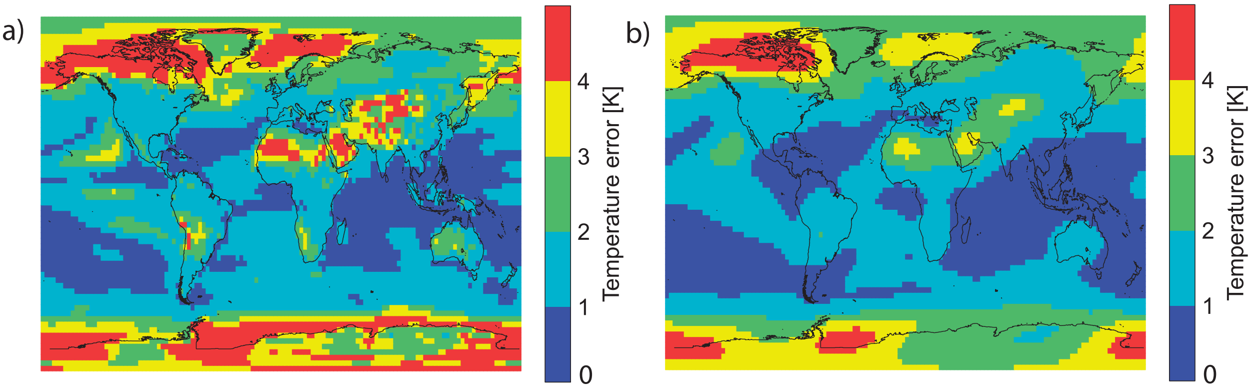

FIG. 1. Effect of (a) weak ($\lambda = 100$ km) and (b) medium ($\lambda = 700$ km) field smoothing on the average magnitude of the temperature error of the CMIP3 ensemble for the period 1980–99. With smoothing, some details are lost but the average error is smaller.

$$\sigma^s_{k,l}(\lambda)^2 = \frac{\sum_{i,j}^{I,J}[V_{i,j} - \overline{V}_{k,l}(\lambda)]^2 w_{i,j}(k,l,\lambda)}{\sum_{i,j}^{I,J} w_{i,j}(k,l,\lambda)}.$$

This characterizes the spatial heterogeneity in a region determined by an area proportional to λ^2. A large σ^s indicates that the value at that location is not very informative about the original spatial details, whereas a small σ^s indicates that most points within a distance of about λ are similar, such that not much of the spatial pattern is lost with smoothing. The parameter λ was sampled logarithmically between 100 and 10 000 km since most variations in the field occur at small scales and tend to decrease toward global scales. For illustration, the typical magnitude of the CMIP3 present-day temperature errors are shown in Fig. 1 after a small ($\lambda = 100$ km) and a medium ($\lambda = 700$ km) smoothing. As an alternative to the Gaussian smoothing, a simple average over grid points is performed using a step function with weight 1 inside a circular region of radius λ and 0 elsewhere. The difference between the Gaussian and the step-function smoothing is examined in the results section.

c. *Measures of uncertainty*

Three sources of uncertainty have been considered: 1) the model error $E_{i,j}(\lambda)$ at location (i, j) after smoothing with a fixed scale parameter λ, defined as the absolute value of the difference between the simulation and a observation-based set; 2) σ^s, defined as the spatial standard deviation of the variable in a given area (see definition above); and 3) the intermodel spread σ^m, defined as the CMIP3 ensemble standard deviation for a grid point or a region. In a first part, the error and σ^s are analyzed for the period 1980–99 (termed present day here). For the error and this part only, we have subtracted the present-day global error from all simulated data to account for the fact that some models have a global warm or cold error that is not obviously related to their ability to simulate patterns. In a second part, the evolution of these quantities over all successive periods of 20 yr between 1960 and 2100 is studied.

d. *Optimal smoothing for present-day simulations*

One possible way to define an optimal spatial scale for present day is to use a penalty function that accounts for both the error and the spatial variation. The optimal spatial scale should consider the error magnitude E and σ^s as two independent components of uncertainty. To give similar relative importance to both quantities, both components are normalized with $E^{\max}_{i,j}$ and $\sigma^{s,\max}_{i,j}$, the local maximum values of $E_{i,j}$ and $\sigma^s_{i,j}$ across all tested spatial scales. We define the global penalty function $U(\lambda)$ as

$$U(\lambda) = \sqrt{\text{global mean}\left\{\left[\frac{E_{i,j}(\lambda)}{E^{\max}_{i,j}}\right]^2 + \left[\frac{\sigma^s_{i,j}(\lambda)}{\sigma^{s,\max}_{i,j}}\right]^2\right\}}.$$

The optimal spatial scale $\lambda_{\text{opt},m}$ for a given model m minimizes $U(\lambda)$. The optimal spatial length for the entire CMIP3 ensemble is obtained by computing the median value λ_{opt} over all $\lambda_{\text{opt},m}$. We interpret this optimal spatial scale as a typical scale on which the model errors E are reasonably small yet a large portion of the spatial signal is preserved (small σ^s indicating that the unsmoothed values of points nearby are similar to the smoothed value in the center of the area). The resulting optimal scale depends on the choice of the normalization and the definition of the penalty function, which are both partly subjective, and as a consequence the results should be interpreted as illustrative (see sections 3b, 4c).

e. Impact of resolution on model error

Larger computational capacities are often justified with the need for higher resolution. It is assumed that a model run at a higher spatial resolution will provide more reliable information than a model run at a lower resolution. We test this assumption by doing a regression of the error against the resolution for different smoothing values. This is particularly interesting for strong smoothing, as it answers whether higher resolution (in addition to resolving smaller features) also improves the simulation of the large-scale pattern. Resolution is defined here as the typical edge length of a grid cell, calculated as the square root of the earth surface after dividing by the number of grid cells.

f. Robustness of climate change signal

To assess the strength of the predicted climate change signal compared to σ^m, the ensemble robustness ratio R is defined:

$$R_{ij}(\lambda) = \left|\frac{\Delta\overline{V}_{i,j}(\lambda)}{\sigma^m_{i,j}(\lambda)}\right|.$$

This ratio is defined as a function of the location (i, j) and λ, with $\Delta\overline{V}$ being either the smoothed field of temperature or precipitation change (based on the multimodel mean value). As proposed by Murphy et al. (2004), the climate change signal is considered robust if the absolute value of the ratio is larger than 2, that is, the predicted climate change signal is at least twice as large as the uncertainty across models.

g. Initial error preservation

Models are often evaluated and calibrated toward an observation-based dataset with the hope that this would ensure skill for a prediction. But do the initial model errors during 1960–79 also explain the errors in the future? Is a good model for the present day still good in the future, and on what spatial scale is this relationship strongest? We examine these questions with the help of a perfect model approach. As the global error of the initial period 1960–79 dominates the future error signal at all spatial scales, we subtract it from the data. This is justified since the focus lies rather on the spatial error pattern generated after 1960–79. The relation between initial and future errors and simulated change and the role of the spatial scale is studied by a squared correlation index $I(t,\lambda)$ as a function of smoothing scale and projection lead time. For a given time period t and λ the following squares ρ^2 of the correlation value are calculated at each grid point and then globally averaged:

$$I^{(1)}_{i,j}(t,\lambda) = \rho^2[\mathbf{E}_{i,j}(1960-1979,\lambda),\mathbf{E}_{i,j}(t,\lambda)] \quad \text{and}$$

$$I^{(2)}_{i,j}(t,\lambda) = \rho^2[\mathbf{E}_{i,j}(1960-1979,\lambda),\mathbf{C}_{i,j}(t,\lambda)],$$

with $\mathbf{E}_{i,j}$ being a vector of the 23 error values for a certain time, smoothing, and grid point (i, j). The length of the vector is 23 and is equal to the number of CMIP3 models, 24, minus one model that serves as reference to calculate the error. This perfect model approach is repeated 24 times so that each model is used as reference once. The 24 possible $I(t, \lambda)$ are then averaged and return a single representative fraction of explained variance for a certain time period and smoothing. Here, $\mathbf{C}_{i,j}$ is the difference between the simulated variable change (temperature or precipitation) of two models for a certain time, smoothing, and grid point. We assume the relations to be linear. The squared correlation coefficient thus equals the fraction of explained variance under the hypothesis that the predicted variable is normally distributed at any predicting value (von Storch and Zwiers 2004, section 8.2.4). This hypothesis is met by the distributions of the projected error magnitudes among the 23 simulations. As the fraction of explained variance is an additive value, $I^{(1)}_{i,j}$ and $I^{(2)}_{i,j}$ can be globally averaged for a given t and λ. The quantity $I^{(1)}$ therefore measures the fraction of future error in the variable that is explained by the initial error during 1960–79 (control error), while $I^{(2)}$ measures the fraction of error in the variable change (i.e., the simulated difference rather than the variable itself) that is explained by the initial error during the reference period 1960–79. In other words, $I^{(1)}$ describes the persistence of the initial errors over time, whereas $I^{(2)}$ describes the relation between initial errors and trend errors. High fractions of explained variance in $I^{(2)}$ indicate that the mean state climate for the present-day period is a good indicator for the model consistency in the future—that is, two models with a similar present-day state will simulate similar changes, and therefore a model close to the present-day climate of the real world would hopefully produce an accurate prediction of the changes of the real world.

3. Results

a. Field smoothing and measures of uncertainty

In a first step, the error and σ^s (absolute values averaged over space) are quantified for present-day simulations at various smoothing scales and for each model. The largest errors compared to the observation-based dataset are found at the smallest tested smoothing (λ = 100 km, essentially equivalent to no smoothing) for all models. As shown in Fig. 2, the errors decrease monotonically and all curves converge to zero as λ increases. Note that the

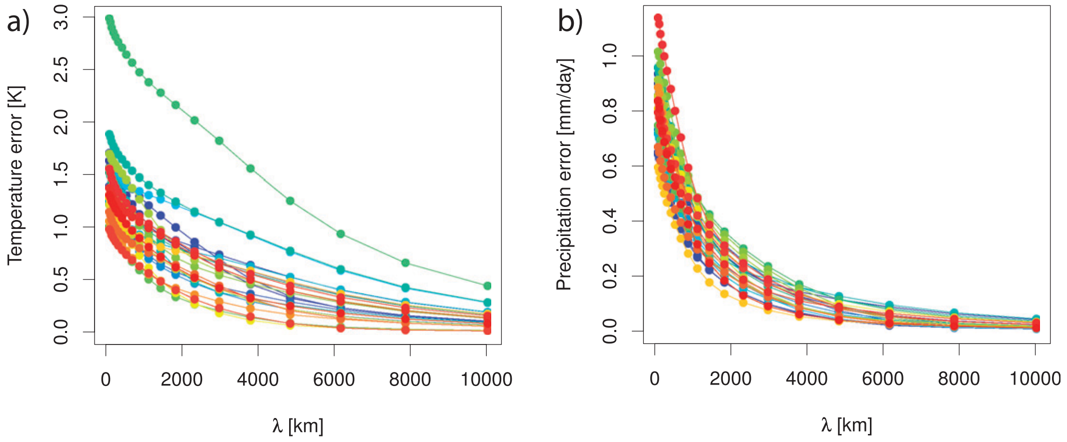

FIG. 2. Global average of the absolute value of the error as function of λ, for each CMIP3 model. (a) Temperature (ERA-40 as reference) and (b) precipitation (CMAP as reference) for the period 1980–99. Smoothing the data reduces the error between simulation and observation.

global error was initially subtracted. For precipitation the error reduces faster than for temperature and even a weak smoothing significantly improves the agreement with observation.

The ranking of the models (from the error point of view) depends on the spatial scale considered. In general, a model performing well on small scales tends to also perform well after smoothing while the opposite is not necessarily true. Because the models get more and more similar with increasing spatial scales, performing well at large scales does not guarantee good agreement with the observation-based dataset at smaller scales. A more detailed study of how such model rankings evolve over time is given later in section 3e. Two observational datasets (CMAP and GPCP) are available in the case of precipitation. If one reference is treated as the true data and the other serves as an additional model, the best performing model at local scales is the alternative observation-based dataset. However, that is not true at large scales. In the case of CMAP being the reference, GPCP is even among the five worst models. Not surprisingly, both datasets differ from all models on small scales because the models are unable to resolve some small-scale patterns, while this does not seem to hold for large scales. Without judging which observation-based dataset is more realistic, this analysis highlights that observational uncertainty in variables other than temperature may be large and should be considered when developing metrics for model evaluation.

In contrast to the error, σ^s is monotonically increasing, as shown in Fig. 3. For the smallest smoothing, σ^s is near

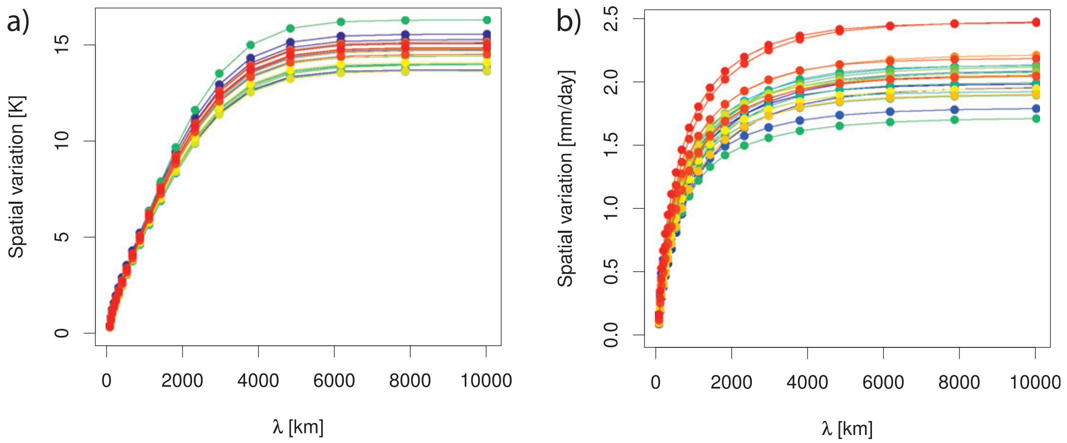

FIG. 3. Global average of σ^s as function of λ for each CMIP3 model and for (a) temperature and (b) precipitation for the period 1980–99. The spatial variation is defined as the spatial standard deviation of the variable within a given spatial area. The larger the spatial scale, the larger the standard deviation of the variable encompassed by the smoothing.

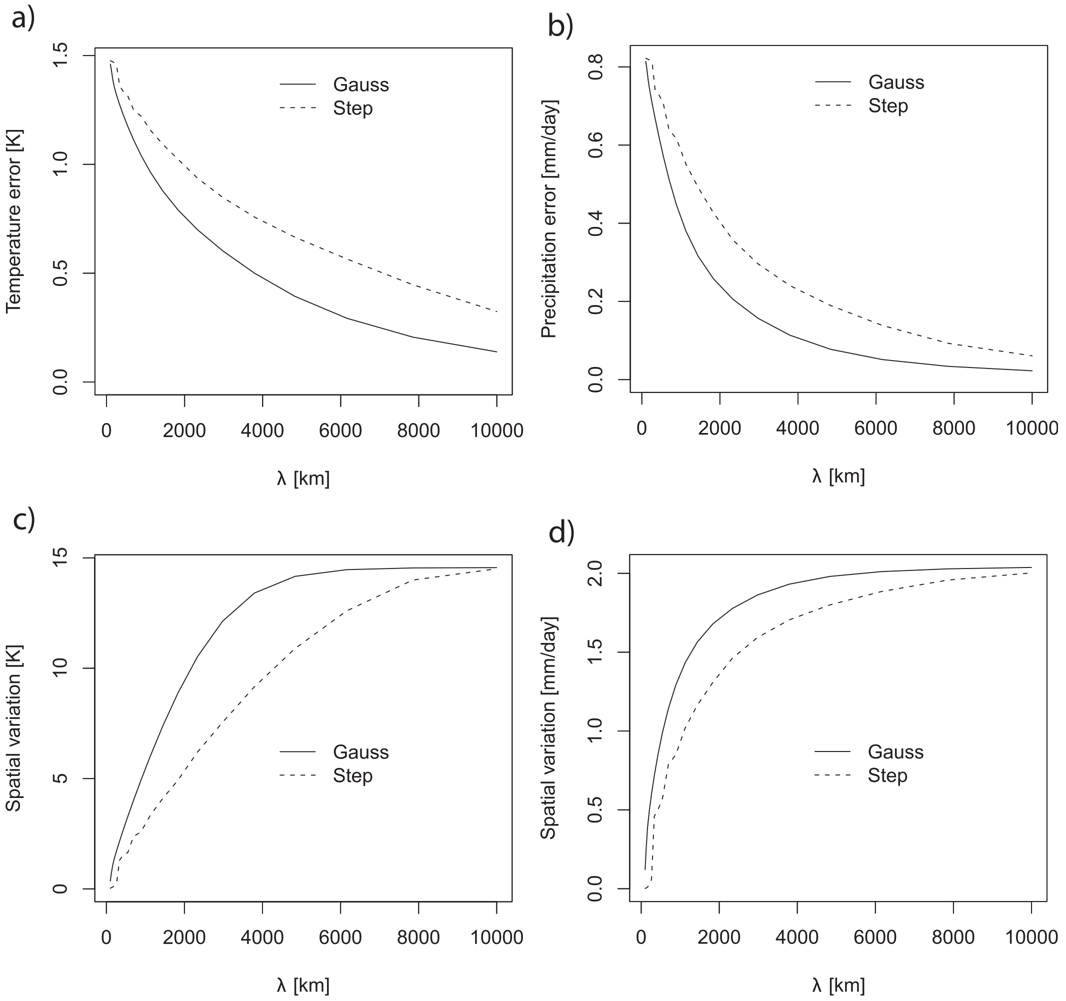

FIG. 4. Gaussian (solid line) vs step-function (dashed line) smoothing for (a),(c) surface temperature and (b),(d) precipitation applied to the typical CMIP3 global average error (a),(b) magnitude and (c),(d) spatial variation.

zero because the standard deviation over all points gives virtually no weight to neighboring points. At large scales, it converges to a constant value representing the standard deviation across all grid points. For temperature data, the curves are approximately constant above 6000 km while for precipitation they stabilize earlier at 4000 km.

Because different smoothing techniques are likely to produce different results, the Gaussian smoothing and a simple average over neighboring grid points using a step function (see section 2b) are compared. Figure 4 shows the typical CMIP3 global average E and σ^s as function of the spatial length. Both techniques return the same values at the grid point and the global spatial scales. The largest difference is the rate of change, which is faster in the case of the Gaussian smoothing. The step-function smoothing shows irregularities between 0 and 1000 km as a hard threshold is more likely to create artifacts when moving across mountain ranges or coastlines. The qualitative behavior, however, is similar.

Figure 5a shows that σ^m (measured as the standard deviation across all models after smoothing and representing model dissimilarities) is larger at local scales and can be reduced with a stronger smoothing, as in the case of the error. At local scales, σ^m over time remains relatively constant. At large scales, however, σ^m is smaller but clearly increases with time. The reason is that global-scale dissimilarities are related to the transient temperature change and evolve with the same magnitude as the global error (Knutti et al. 2008). The local dissimilarities, however, are less related to global warming and dominated by model errors, and thus almost time independent. The reduction of σ^m for precipitation occurs more rapidly than for temperature, similar to the case of the error (see Fig. 5b). In contrast to temperature, the precipitation

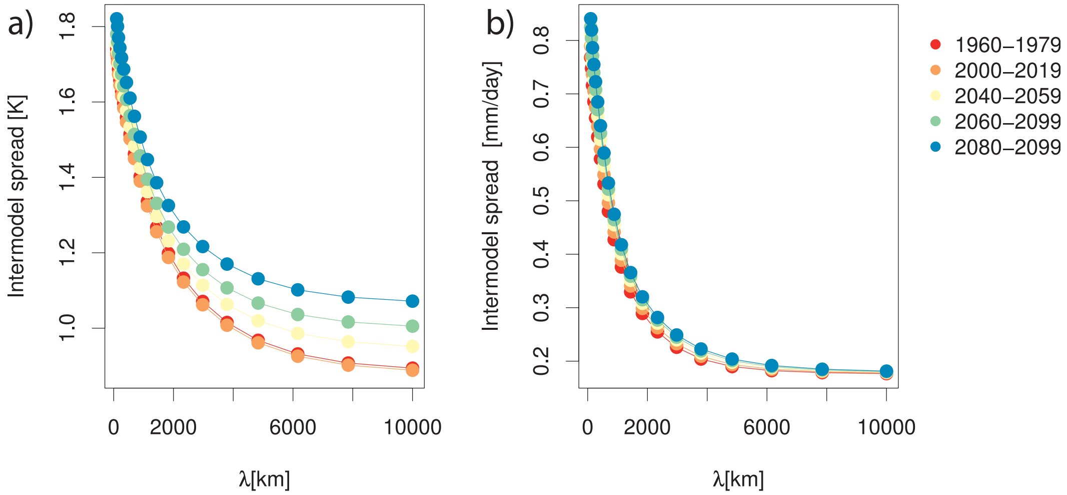

FIG. 5. Intermodel spread representing the model dissimilarities, globally averaged and as a function of λ for (a) temperature and (b) precipitation for different time periods from 1960 to 2080. Models tend to show smaller dissimilarities at larger spatial scales.

σ^m does not change much over time at any of the tested scales because the precipitation trends are rather small compared to the model dissimilarities.

b. *Optimal smoothing for present-day simulations*

The optimal smoothing that minimizes the root-mean-square value of $U_{i,j}(\lambda)$ indicates an approximate scale where the error is reasonably small through spatial averaging, yet most of the local climate pattern is preserved. The results for different regions and smoothing techniques range between 185 and 2080 km and are shown in Table 1. The step-function smoothing produces optimal scales about 2 times larger than the Gaussian smoothing but the results are qualitatively similar. The reason is that the step function eliminates all influences from grid points farther away than λ whereas the Gaussian gives nonzero weight even to very remote points. In general, the temperature field allows a larger smoothing than the precipitation field because the spatial variation for temperature is smaller. This is even more pronounced in the tropics for convective precipitation, where the spatial variation is much larger than the error magnitude, leading to an optimal smoothing close to the gridpoint scale. As the choice of the penalty function is partly subjective, we have investigated two other definitions. For example, if the error and the spatial variation are simply added without normalization, σ^s quickly dominates E and the gridpoint scale is the optimal choice. Different applications may require different weighting in the penalty function. Rather than defending any particular choice of a penalty function, the idea here is to demonstrate the two opposing trends of model error and spatial variation. Trying to minimize both of these components implies a typical length scale over which the model results should be aggregated; that the length scale is larger for temperature than for precipitation globally, is larger for temperature in the tropics, is larger for precipitation in the extratropics, and is larger for temperature over ocean than over land. These general conclusions should be robust against different definitions of the penalty function, provided that an optimal scale exists.

c. *Impact of the resolution on model error*

The correlation between error and model resolution (i.e., the original resolution at which the model is run, not the smoothing scale) is an indication of the benefit of higher resolution in representing current climate. In the case of precipitation, correlations between error and resolution were lower than 0.5 and the regression slopes were never statistically significant using the F test with a 0.05 significance level, thus no clear relation seems to exist at least within the relatively narrow range of resolutions covered by CMIP3. In Fig. 6a, a scatterplot of the relation is shown for temperature at the gridpoint scale ($\lambda = 100$ km),

TABLE 1. Optimal smoothing length (km) for surface air temperature (TAS) and precipitation (PR) for various regions using the Gaussian or the step-function smoothing.

	Gaussian		Step function	
Region	TAS	PR	TAS	PR
Global	695	207	1280	428
Tropics	1130	162	2340	428
Extratropics	428	1130	886	2080
Land only	428	207	886	428
Ocean only	886	185	1440	428

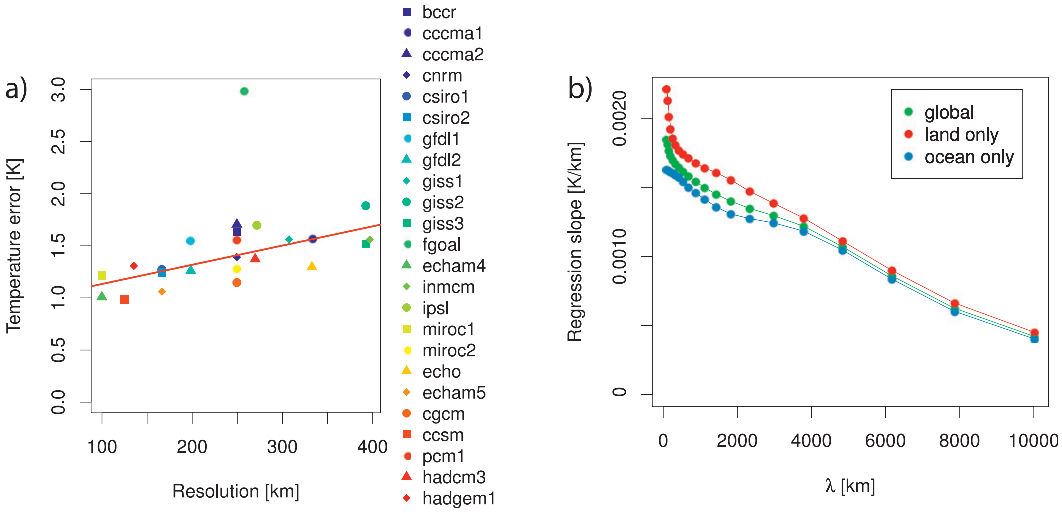

FIG. 6. (a) Relation between absolute value of the temperature error (globally averaged) at the gridpoint scale (λ = 100 km) and native model resolution. The linear regression is indicated by a red line (with the FGOALS model excluded). (b) Regression slope vs λ for land or ocean only and the entire globe. Whereas a higher resolution improves performance at local scales, its impact at large scales is limited. In the case of precipitation, no relation between resolution and error could be established.

considering the whole globe (land and oceans). At this scale, the error is reduced from about 1.5 to 1 K from the coarsest to the highest resolution. The slope of the regression line characterizes the strength of the relation between the resolution and the error and is calculated for all spatial scales, global, land, and oceans. The Flexible Global Ocean–Atmosphere–Land System Model (FGOALS) model is a clear outlier and is excluded for this part of the analysis. The linear regression slopes are displayed in Fig. 6b and are always statistically significant using the same test as before. Not surprisingly, the benefit of high resolution is largest at the smallest scales and over land, where the topography is more complex and higher resolution can probably resolve more local processes. However, the benefit of the resolution quickly decreases with smoothing approaching 500 km. At scales above about 500 km, the slope dependence on λ is similar in all cases.

d. Robustness of climate change signal

We define R as the absolute value of the ratio of the climate change to σ^m. The geographic distribution of the robustness at the gridpoint scale (λ = 100 km) is first shown in the maps in Fig. 7 for the end of the century as an example. A striking but well known feature is that temperature changes are clearly more robust than precipitation (Räisänen 2001). While temperature robustness is especially weak in the North Atlantic and over the

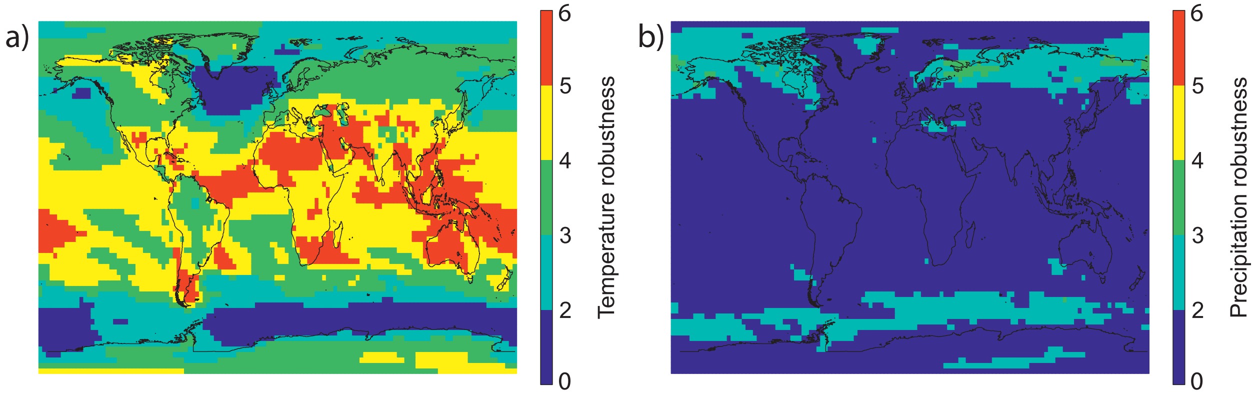

FIG. 7. Maps of R for the 2080–99 climate change signal since 1960–79 at the gridpoint scale (λ = 100 km) for (a) temperature and (b) precipitation. The climate change signal is said to be robust if R is larger than 2.

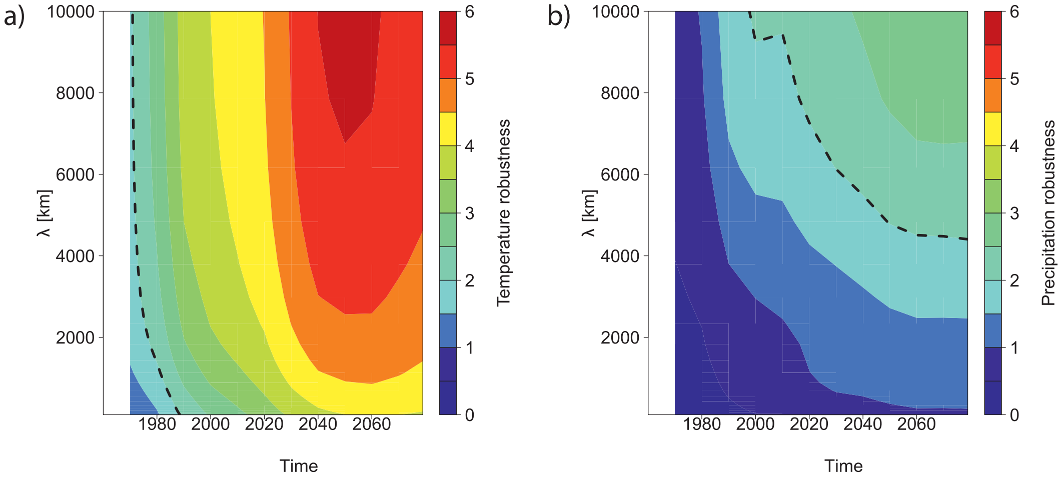

FIG. 8. Profiles of R for the climate change signal since 1960–79 (globally averaged) for (a) temperature and (b) precipitation. Here, R is defined to be robust above the dashed line where the signal is twice as large as the standard deviation characterizing the intermodel spread. While temperature simulations are mostly robust everywhere, precipitation only gets robust later and for continental or large scales.

Southern Ocean where some models indicate cooling while most show warming, the models agree for the rest of the globe. Precipitation simulations show good agreement over high latitudes and the Mediterranean Sea. The temporal and spatial behavior of R is depicted in Fig. 8. Not unexpected and in agreement with detection/attribution and future projection studies (Barnett et al. 2005; Meehl et al. 2007b), the temperature signal is robust at all scales after a few decades and clearly exceeds σ^m. In contrast, simulated precipitation changes agree about 50 years later and on continental scales only, both because of larger model differences and large interannual variability.

e. Initial error preservation

The global average of the explained variance fractions between initial and future model errors $I^{(1)}(t, \lambda)$ is shown in a contour plot in Fig. 9 as function of time and spatial scale. The time axis is divided into 12 intervals between 1960 and 2099. For both temperature and precipitation the fraction of explained variance is high at

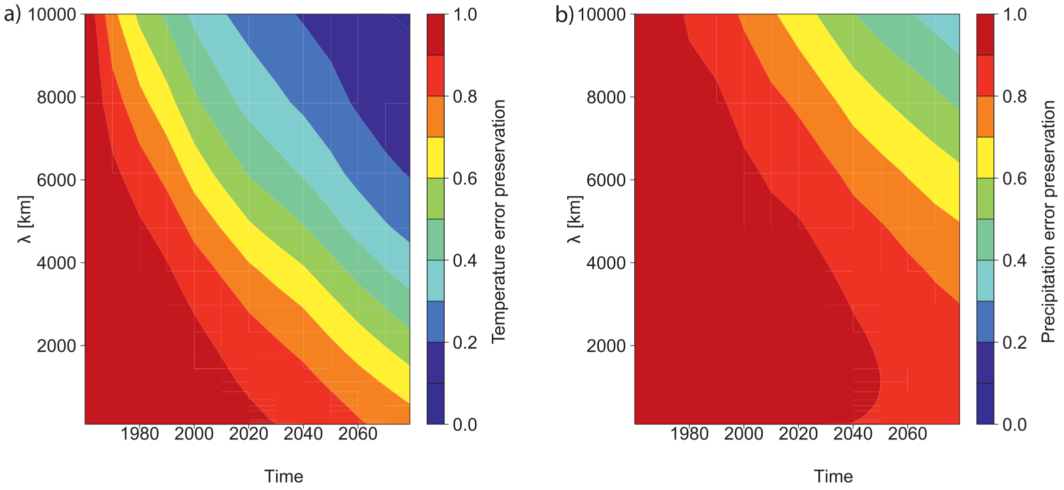

FIG. 9. Preservation of the initial errors among the models through time and spatial scales, defined as the fraction of variance of the future errors that is explained by the initial model errors in 1960–79 (globally averaged). The time axis is divided into 12 intervals between 1960 and 2099. The perfect model approach was used here for (a) temperature and (b) precipitation. Values near 1 indicate that differences between models strongly persist over time, while values near 0 indicate no relation between differences at the beginning and the end of the simulation.

local scales until the end of the twenty-first century. At that time and at larger scales, however, the fraction of explained variance vanishes. The decay of the fraction of explained variance with time is slower at local scales than at large scales. In agreement with Giorgi and Mearns (2002) and Räisänen (2007), a linear regression between initial and future model errors for scales and periods exhibiting a high correlation yield regression slopes very close to one, supporting the conclusions that initial differences between models are well preserved over time, in particular on small scales. The relation $I^{(2)}(t, \lambda)$ between initial model differences and trend errors was also analyzed. In contrast to the case before, only very small correlations are seen for either temperature or precipitation (not shown). Therefore, the initial error seems to relate only weakly to simulated future trends. The interpretation of these results is given in section 4.

4. Discussion

a. Data

Regarding the data, we have used the A1B scenario for the analysis of future projections, but the choice of a specific scenario should not impact the presented results because the ratio of temperature change to forcing is approximately constant across scenarios (Knutti et al. 2008; Gregory and Forster 2008) and simulated patterns tend to be similar for all scenarios (e.g., Meehl et al. 2005; Washington et al. 2009; Meehl et al. 2007b). The number of runs available for each model varies from one to nine but very few models have more than five runs. We selected arbitrarily the first run in the list. This choice is not critical since internal variability after averaging the data on 20 yr is relatively small in comparison to intermodel differences. Choosing periods of 20 yr is a compromise between having enough data inside a period to avoid too much internal variability and having enough periods to cover the observation period between 1960 and 2000. Averaging initial condition ensemble members only for some of the models would inappropriately reduce variability for some models. The model evaluation was only done for temperature and precipitation, but of course other important fields such as sea level pressure would be important to consider for a more complete study.

b. Measures of uncertainty

The magnitude of the error rather than error itself has been chosen for the analysis, because part of the multimodel error gets canceled in a multimodel mean (Räisänen 2007) and the tendency of the error magnitude to decrease with increasing smoothing is masked. Our results suggest that smoothing the data before the model evaluation reduces the disagreement between the models and an observational dataset, in particular in the case of a variable that is difficult to simulate and locally heterogeneous, such as precipitation. This does not ignore the weaknesses of the models but may help to focus on features that are relevant at a certain spatial scale. Two datasets were used to evaluate precipitation, and it is interesting that on larger scales the similarity between some models and one dataset is larger than the similarity of the two datasets. For small scales, the two datasets are similar and all models are different because the models all share similar limitations in parameterizing or discretizing physical processes (Tebaldi and Knutti 2007; Knutti et al. 2010b) and are unable to correctly capture certain small-scale patterns in precipitation.

As model errors decrease when results are aggregated on larger scales, grid points from more climatic regimes are averaged together and the uncertainty related to σ^s increases. The information provided by the models gets blurred and the simulation signal is less precise than at the gridpoint scale. Similar to the error, σ^m for precipitation is reduced faster by smoothing than for temperature. This is because precipitation differences are dominated by small-scale features whereas temperature differences vary at larger spatial scales. Although σ^m for temperature is lowest at large scales, it increases faster with time at large scales owing to the different transient warming rates of the models. The case of precipitation is interesting because σ^m is almost constant in time at each tested spatial scale. The reason is that precipitation changes are only on the order of a few percent and can be of opposite sign in nearby areas and therefore get partly averaged out at larger spatial scales. On the other hand, simulated precipitation in the baseline climate can easily vary by a factor of two locally, so σ^m is dominated by the climatological errors at all times.

c. Optimal smoothing for present-day simulations

Combining the error and σ^s into an overall penalty function where the error is substantially reduced yet the main spatial patterns are still preserved is possible but the results depend on the definition adopted. Our results suggest that there are different optimal scales, depending on which variable is analyzed as well as which region is considered. From a numerical perspective several grid points are needed to discretize the partial differential equations governing the climate models. Therefore, interpreting scales smaller than at least several grid points is dangerous because of numerical instabilities and errors generated. It is interesting to compare the optimal scales obtained here with the choices made in publications that aggregate regional climate change (e.g., Giorgi and Francisco 2000; Tebaldi et al. 2005; Christensen et al. 2007; Mahlstein and Knutti 2009). Many of these studies provided regional averages over 26 land regions. If we divide the total land area by 26 and take the square root of that (corresponding to the length and width of a region if the land was divided into 26 regions of equal area), we find characteristic length scales of about 2400 km, larger than those obtained here with the Gaussian smoothing but in closer agreement with the step-function scales. We argue that climate scientists may in fact often choose regions and scales based on a similar informal optimization procedure, trying to maximize the regional detail (e.g., for impact studies). But knowing that errors and σ^m are largest at local scales, they aggregate results into regions encompassing multiple grid points. Of course, other aspects such as the communication of the results play an important role, and the optimal spatial scales found in this analysis should be seen as an approximate estimate that depends on the location, the variable, the temporal and spatial variability, as well as the uncertainty in the observation. Different definitions of a penalty function and an optimal spatial scale are possible, and the one chosen here should be interpreted as an illustration that provides insight into how different quantities depend on the spatial scale, rather than as a definitive answer.

d. *Impact of resolution on model error*

The resolution of models is correlated with their performance in simulating temperature but apparently not precipitation (at least in the CMIP3 ensemble). This might be because resolving some processes related to clouds and precipitation would require much higher resolution than any of the global models currently have.

Alternatively, it might also be that higher-resolution models produce precipitation with higher geographical variability. Since the precipitation field is more variable than temperature, a simple shift in precipitation pattern could penalize the model performance. High-resolution models have the clearest advantage in reproducing current temperature over land, on scales between local and 500 km, partly because of a better representation of the topography. The globally averaged error is reduced from about 1.5 to 1 K from the coarsest to the highest resolved models at the smallest scale, where the relation is strongest. However, given the cost of higher resolution, the benefit may be seen as rather small. As already noted by Santer et al. (2009), the Canadian Centre for Climate Modelling and Analysis (CCCma) Coupled General Circulation Model, version 3.1 (CGCM3.1) and the Japanese Model for Interdisciplinary Research on Climate version 3.2 (MIROC3.2) were both run at higher- and lower-resolution configurations, but despite the use of higher resolution the error was not much reduced. For some variables, a higher resolution may eliminate a parameterization and allow direct dynamical computations instead. In the ocean, for example, a lot of energy is contained in small-scale eddies. Therefore, the relation between resolution and sea surface temperature might be stronger than for surface air temperature. In general, it is not easy to separate the effects of higher resolution and a more comprehensive representation of processes. The groups running models at highest resolution are often also those with the longest experience in building models and with the largest number of people developing the model (of course resolution can be changed, but each model has one or a few standard resolutions that are commonly used and for which it has been optimized). So resolution, rather than just a numerical property, should probably be seen more as an indicator of overall sophistication, effort, computing power, and financial resources going in a model.

e. *Robustness of climate change signal*

Maximizing R is desirable and is likely to improve with newer climate models generations. A closer look at Fig. 8b for precipitation compared to Fig. 10.12 of Meehl et al. (2007b) summarizes the benefit of this study: if models are compared at the gridpoint scale they do not agree in their trends over large areas on the globe, whereas our study shows that they do but only for regions with 4000 km as typical scale and trends beyond the period 2050–69. Trends are relatively weak and local precipitation is difficult to simulate, therefore larger regions are needed to detect the precipitation change signal and simulate robust trends, consistent with the results of the precipitation attribution study by Zhang et al. (2007). Note that even if the global average of the robustness is below a given threshold, there are of course regions where the robustness is high. For example, the simulated increase in precipitation in high northern latitudes is significant and robust even at small scales and in the near future.

f. *Initial error preservation*

The climatological errors in mean temperature and precipitation in temperature and precipitation in the CMIP3 models are surprisingly constant over time, in particular on small scales. This error preservation vanishes toward the end of the century at large scales, indicating that differences in climate change have a larger scale than the differences in error magnitudes. This is consistent with the fact that climatological errors and simulated changes are weakly correlated (Murphy et al. 2004; Knutti et al. 2006, 2010b). Local errors are more likely to be the result of deficiencies in simulating particular processes that are important locally (e.g., not resolving a mountain range), and these are usually more persistent over time. The fraction of explained variance between initial and future errors is larger for precipitation than for temperature, which may seem surprising at first since precipitation is more difficult to simulate. The reason for this persistence is that changes in precipitation are quite small in many regions compared to the control errors. Therefore, future errors are essentially a sum of initial errors plus some trend, with the former dominating the latter. The result is that two models having similar errors in their initial state will tend to have similar errors in the future at the same location.

The lack of correlation between the initial errors and future trends errors is rather disturbing. It is commonly assumed that models with small initial error are more accurate in predicting future trends (e.g., IPCC AR4 Frequently Asked Questions (FAQ) 8.1, Randall et al. 2007]. But reality suggests otherwise, as shown by the lack of correlation between past or future predicted warming with present-day simulated temperatures (Tebaldi and Knutti 2007; Jun et al. 2008; Knutti et al. 2010b; Weigel et al. 2010). As a consequence, knowing the discrepancy between present-day simulations and observations of the mean climate state does not immediately help to constrain the estimation of future trend error. On the other hand, there is clear evidence and physical reasons for a relation of past greenhouse gas attributable warming and future warming (e.g., Stott and Kettleborough 2002). Clear relations also exist for local processes, for example, a correlation between past and future sea ice loss in the Arctic (Boé et al. 2009). These points are important to keep in mind when weighting the models for the future projections, with weights based on performance in the past (Knutti et al. 2010a; Knutti 2010).

5. Conclusions

Although small spatial scales are most important to determine specific climate impacts, this is precisely the scale where climate is most difficult to simulate and where model errors and intermodel spread are largest. Climate scientists therefore often aggregate data to regions where the above problems are less severe, even though some spatial information gets lost in those processes. Here, we have done this in a formal way and have demonstrated how the spatial variation σ^s, the model spread σ^m, the model error in climatology, and the persistence of errors depend on the spatial scale of averaging. We have shown that the error and the intermodel spread can be significantly reduced by smoothing the data (consistent with earlier results by Räisänen 2001), however, at the price of losing spatial detail (expressed in our case as an increase in the spatial spread).

Our results support typical scales between the gridpoint and 2000 km depending on the variable, the location, and the smoothing technique. Although there are of course various definitions of an optimal scale for different problems, we suggest that some form of spatial aggregation should be considered to provide a more robust estimate of climate change.

Our analysis also reveals that model resolution in CMIP3 seems to only affect performance in simulating present-day temperature for small scales over land. Results may differ for other quantities, but given the limited advantages of high resolution even for reproducing present-day climate, we speculate that pushing model resolution alone is unlikely to improve future predictions and reduce uncertainties, unless a more complete understanding of the physical and biogeochemical process is incorporated and the models are recalibrated. This is consistent with the fact that uncertainties in climate projections have not decreased significantly in the last decade despite massive computational advances allowing for higher model resolution. As a consequence, regional high-resolution models and downscaling may provide greater spatial detail but not necessarily higher confidence in local projections to determine the impacts of climate change.

Finally, we have shown that the initial model errors of 1960–79 persist over time in particular at small spatial scales, justifying to some extent the common practice of focusing on anomalies from a control simulation rather than absolute values (although that is unlikely to work well for more complicated quantities, see Buser et al. 2009). In agreement with earlier studies, there is a lack of correlation between straightforward measures of initial errors and future trends. We have shown the difficulty in relating model skill based on present-day climatological errors (as characterized by Reichler and Kim 2008) to prediction skill in the future, whatever spatial scale is chosen.

Acknowledgments. We acknowledge the international modeling groups for providing their data for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model data, the WCRP/CLIVAR Working Group on Coupled Models (WGCM), and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panels for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. The CMIP3 Multi-Model Data Archive at Lawrence Livermore National Laboratory is supported by the Office of Science, U.S. Department of Energy.

REFERENCES

Adler, R. F., and Coauthors, 2003: The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeor.,* **4,** 1147–1167.

Barnett, T., and Coauthors, 2005: Detecting and attributing external influences on the climate system: A review of recent advances. *J. Climate,* **18,** 1291–1314.

Boé, J. L., A. Hall, and X. Qu, 2009: September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nat. Geosci.,* **2,** 341–343.

Buser, C., H. Künsch, D. Lüthi, M. Wild, and C. Schär, 2009: Bayesian multi-model projection of climate: Bias assumptions and interannual variability. *Climate Dyn.,* **33,** 849–868, doi:10.1007/s00382-009-0588-6.

Christensen, J., and Coauthors, 2007: Regional climate projections. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 847–940.

Giorgi, F., and R. Francisco, 2000: Uncertainties in regional climate change prediction: A regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. *Climate Dyn.,* **16,** 169–182.

——, and L. O. Mearns, 2002: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the Reliability Ensemble Averaging (REA) method. *J. Climate,* **15,** 1141–1158.

——, and ——, 2003: Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophys. Res. Lett.,* **30,** 1629, doi:10.1029/2003GL017130.

Gregory, J. M., and P. M. Forster, 2008: Transient climate response estimated from radiative forcing and observed temperature change. *J. Geophys. Res.,* **113,** D23105, doi:10.1029/2008JD010405.

Jun, M., R. Knutti, and D. W. Nychka, 2008: Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Stat. Assoc.,* **103,** 934–947.

Knutti, R., 2008a: Should we believe model predictions of future climate change? *Philos. Trans. Roy. Soc.,* **A366,** 4647–4664.

——, 2008b: Why are climate models reproducing the observed global surface warming so well? *Geophys. Res. Lett.,* **35,** L18704, doi:10.1029/2008GL034932.

——, 2010: The end of model democracy? *Climatic Change,* **102,** 395–404, doi:10.1007/s10584-010-9800-2.

——, G. A. Meehl, M. R. Allen, and D. A. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Climate,* **19,** 4224–4233.

——, and Coauthors, 2008: A review of uncertainties in global temperature projections over the twenty-first century. *J. Climate,* **21,** 2651–2663.

——, G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and L. Mearns, 2010a: Good practice guidance paper on assessing and combining multimodel climate projections. Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections, Stocker et al., Eds., IPCC Working Group I Technical Support Unit, 13 pp. [Available online at https://www.ipcc-wg1.unibe.ch/guidancepaper/IPCC_EM_MME_GoodPracticeGuidancePaper.pdf.]

——, R. Furrer, C. Tebaldi, and J. Cermak, 2010b: Challenges in combining projections from multiple climate models. *J. Climate,* **23,** 2739–2758.

Mahlstein, I., and R. Knutti, 2009: Regional climate change patterns identified by cluster analysis. *Climate Dyn.,* **35,** 587–600, doi:10.1007/s00382-009-0654-0.

Meehl, G. A., W. M. Washington, W. D. Collins, J. M. Arblaster, A. X. Hu, L. E. Buja, W. G. Strand, and H. Y. Teng, 2005: How much more global warming and sea level rise? *Science,* **307,** 1769–1772.

——, C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007a: The WCRP CMIP3 multimodel dataset—A new era in climate change research. *Bull. Amer. Meteor. Soc.,* **88,** 1383–1394.

——, and Coauthors, 2007b: Global climate projections. *Climate change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 747–785.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, and M. Collins, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature,* **430,** 768–772.

Nakicenovic, N., and Coauthors, 2000: *IPCC Special Report on Emissions Scenarios*. Cambridge University Press, 570 pp.

Räisänen, J., 2001: CO$_2$-induced climate change in CMIP2 experiments: Quantification of agreement and role of internal variability. *J. Climate,* **14,** 2088–2104.

——, 2007: How reliable are climate models? *Tellus,* **59A,** 2–29.

Randall, D., and Coauthors, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 589–662.

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.,* **89,** 303–311.

Santer, B. D., and Coauthors, 2009: Incorporating model quality information in climate change detection and attribution studies. *Proc. Natl. Acad. Sci. USA,* **106,** 14 778–14 783.

Stainforth, D. A., M. R. Allen, E. R. Tredger, and L. A. Smith, 2007: Confidence, uncertainty and decision-support relevance in climate predictions. *Philos. Trans. Roy. Soc.,* **A365,** 2145–2161.

Stott, P. A., and J. A. Kettleborough, 2002: Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature,* **416,** 723–726.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc.,* **A365,** 2053–2075.

——, R. L. Smith, D. Nychka, and L. O. Mearns, 2005: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. Climate,* **18,** 1524–1540.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

von Storch, H., and F. Zwiers, 2004: *Statistical Analysis in Climate Research*. Cambridge University Press, 485 pp.

Washington, W. M., R. Knutti, G. A. Meehl, H. Y. Teng, C. Tebaldi, D. Lawrence, L. Buja, and W. G. Strand, 2009: How much climate change can be avoided by mitigation? *Geophys. Res. Lett.,* **36,** L08703, doi:10.1029/2008GL037074.

Weigel, A., R. Knutti, M. Liniger, and C. Appenzeller, 2010: Risks of model weighting in multimodel climate projections. *J. Climate,* **23,** 4175–4191.

Xie, P. P., and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.,* **78,** 2539–2558.

Zhang, X. B., F. W. Zwiers, G. C. Hegerl, F. H. Lambert, N. P. Gillett, S. Solomon, P. A. Stott, and T. Nozawa, 2007: Detection of human influence on twentieth-century precipitation trends. *Nature,* **448,** 461–466.