OPEN ACCESS
IOP PUBLISHING ENVIRONMENTAL RESEARCH LETTERS
Environ. Res. Lett. **8** (2013) 024028 (9pp) doi:10.1088/1748-9326/8/2/024028

Sensitivity of inferred climate model skill to evaluation decisions: a case study using CMIP5 evapotranspiration

Christopher R Schwalm[1], Deborah N Huntinzger[1,2], Anna M Michalak[3], Joshua B Fisher[4], John S Kimball[5], Brigitte Mueller[6], Ke Zhang[7] and Yongqiang Zhang[8]

[1] School of Earth Sciences and Environmental Sustainability, Northern Arizona University, Flagstaff, AZ 86011, USA
[2] Department of Civil Engineering, Construction Management, and Environmental Engineering, Northern Arizona University, Flagstaff, AZ 86011, USA
[3] Department of Global Ecology, Carnegie Institution for Science, Stanford, CA 94305, USA
[4] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA
[5] Flathead Lake Biological Station, Division of Biological Sciences, The University of Montana, Polson, MT 59860-6815, USA
[6] Institute for Atmospheric and Climate Science, ETH Zürich, 8092 Zürich, Switzerland
[7] Department of Organismic & Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[8] CSIRO Land and Water, Canberra, ACT, Australia

E-mail: christopher.schwalm@nau.edu

Received 4 February 2013
Accepted for publication 9 May 2013
Published 23 May 2013
Online at stacks.iop.org/ERL/8/024028

Abstract
Confrontation of climate models with observationally-based reference datasets is widespread and integral to model development. These comparisons yield skill metrics quantifying the mismatch between simulated and reference values and also involve analyst choices, or meta-parameters, in structuring the analysis. Here, we systematically vary five such meta-parameters (reference dataset, spatial resolution, regridding approach, land mask, and time period) in evaluating evapotranspiration (ET) from eight CMIP5 models in a factorial design that yields 68 700 intercomparisons. The results show that while model–data comparisons can provide some feedback on overall model performance, model ranks are ambiguous and inferred model skill and rank are highly sensitive to the choice of meta-parameters for all models. This suggests that model skill and rank are best represented probabilistically rather than as scalar values. For this case study, the choice of reference dataset is found to have a dominant influence on inferred model skill, even larger than the choice of model itself. This is primarily due to large differences between reference datasets, indicating that further work in developing a community-accepted standard ET reference dataset is crucial in order to decrease ambiguity in model skill.

Keywords: climate models, model validation, evapotranspiration, CMIP5

S Online supplementary data available from stacks.iop.org/ERL/8/024028/mmedia

1. Introduction

A central challenge in the 21st century is to understand and forecast the impacts of global climate change on terrestrial

ecosystems. Numerous advances in understanding the climate system have been driven by model intercomparison projects (e.g., Friedlingstein *et al* 2006; Meehl *et al* 2007; Schwalm *et al* 2010; Taylor *et al* 2012), with confidence in model projections ultimately linked to how well climate models replicate known past features of the climate system (Luo *et al* 2012, Randall *et al* 2007).

The process of systematically reconciling observationally-driven references with climate model output fields, termed benchmarking (Luo *et al* 2012), allows for the quantification of simulation–reference mismatch and ultimately improvements in model formulation (Luo *et al* 2012, Schwalm *et al* 2010). At a minimum, benchmarking requires a skill metric that quantifies the 'distance' between reference and simulated values. More comprehensive benchmarking frameworks track model skill over successive versions of a given model (Gleckler *et al* 2008) and allow for a quantitative evaluation of model skill across multiple fields and models (Randerson *et al* 2009). While benchmarking as a conceptual framework in model evaluation is actively evolving and therefore can be implemented in alternate ways (Abramowitz 2012), we define benchmarking in this study as a systemic framework for confronting simulations with observationally-based and independently-derived reference products similarly scaled to simulation outputs in space and time. This is distinct from other frameworks that confront simulated values with results from statistical or physical models (e.g., Abramowitz 2005, 2012).

Since their initial development, climate models have been routinely compared to observationally-driven references but with little consideration of how the choice of meta-parameters in model evaluation influences inferred model skill (Gleckler *et al* 2008, Jiménez *et al* 2011). Meta-parameters are used here to describe analyst choices (e.g., reference dataset, spatial resolution, regridding algorithm, land mask, time period) that impact simulation–reference mismatch and therefore inferred model skill (see section 2). To improve benchmarking efforts, there is a need to understand how the choice of reference product and other benchmarking meta-parameters influence model skill.

Here we quantify the degree to which inferred climate model skill for a given variable, evapotranspiration (ET), is sensitive to the choice of benchmarking meta-parameters. We do not, strictly speaking, evaluate climate models against ET. Rather, our focus is on assessing how analyst choices impact inferred model skill. Various model types (e.g., climate models, offline land surface models) and reference products (e.g., gross primary productivity, net radiation) are amenable to this goal. This study presents a case study using climate models and ET to illustrate the interdependency between analyst choices and inferred skill. We focus on ET due to the tight coupling of terrestrial water, energy and carbon cycles, the importance of longer-term trends in the hydrological cycle in modulating land sink variability (Schwalm *et al* 2011), and the existence of multiple observationally-based ET references (e.g., Jiménez *et al* 2011; Mueller *et al* 2011; Vinukollu *et al* 2011). Furthermore, these ET reference products are global, potentially tightly-constrained (Vinukollu *et al* 2011), multi-year, and most importantly, are analogous to climate model output both in spatial and temporal scale. We explore the consequences of analyst choice, with emphasis on reference dataset, on inferred individual model skill and rank in simulating ET.

2. Data and methods

We compare six different reference ET products (supplementary table 1 available at stacks.iop.org/ERL/8/024028/mmedia) to simulated ET from eight coupled carbon–climate models (supplementary table 2 available at stacks.iop.org/ERL/8/024028/mmedia) participating in the Coupled Model Intercomparison Project phase 5 (CMIP5) (Taylor *et al* 2012) and using the Earth System Model historical natural experiment (esmHistorical). CMIP5 output is chosen because of its availability and use in the IPCC AR5 framework, as well as its widespread application in climate impact studies. The esmHistorical CMIP5 experiment is selected due to its focus on simulating and evaluating historical conditions (Taylor *et al* 2012). For six of the eight CMIP5 models, only a single esmHistorical realization is available; for those two models with multiple realizations only the first is used.

Of the six ET reference products there is no clear standard. Despite some regional agreement (Mueller *et al* 2011) and consistency with ground measurements (Fisher *et al* 2008, Jung *et al* 2011, Vinukollu *et al* 2011), the gridded ET reference products show disagreement in global annual ET flux (supplementary table 1), with large cross-product variability (Mueller *et al* 2011) and associated differences in latitudinal gradients and seasonal cycles (figure 1). This absence of convergence on a single 'best' ET product stems from the absence of a conclusive ET product intercomparison, though efforts are underway to resolve this (e.g., GEWEX LandFlux/LandFlux-EVAL (Mueller *et al* 2011)). Nonetheless, this lack of benchmark dataset consensus allows us to assess the impact of reference dataset selection on model evaluation.

In addition to varying the choice in ET reference product, we systematically vary: (1) spatial resolution (all model/reference grids as well as uniform 1° and 5° grids); (2) regridding algorithm (nearest neighbor, bi-linear interpolation, and box averaging); (3) land-water mask (all possible combinations of two land cover maps; either IGBP (Loveland *et al* 2001) or SYNMAP (Jung *et al* 2006); and three different per cent land-cover cutoffs for defining land cells); and (4) ten-year analysis period (all possible ten-year periods from 1980 to 2005). All values for each meta-parameter are given in supplementary table 3 (available at stacks.iop.org/ERL/8/024028/mmedia). The result is 68 700 individual model–reference benchmarking experiments (approximately 8500 for each CMIP5 model) based on all possible combinations of meta-parameter and CMIP5 model. In each experiment model simulations and references are translated to a common target grid and land mask with the chosen regridding algorithm (supplementary table 3). Each experiment represents one model evaluation scenario, i.e., a combination of analyst choices. Collectively,

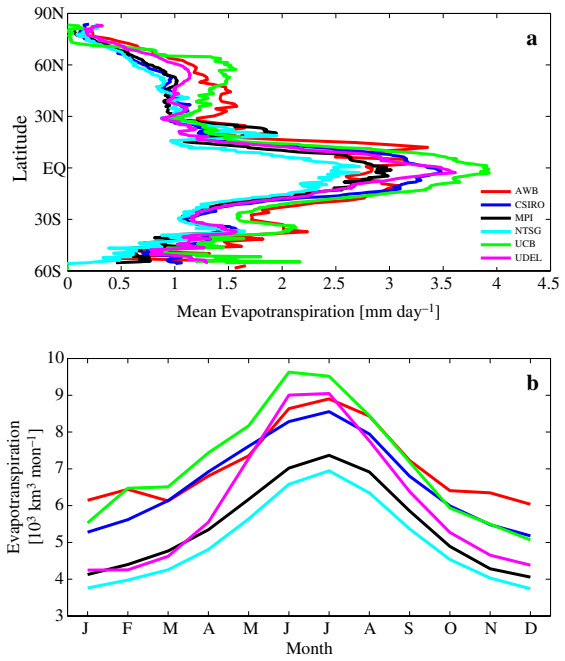

Figure 1. Spatial and temporal patterns in ET. Reference product ET displayed as (a) latitudinal gradients; and (b) a mean seasonal cycle. Values reference land surface excluding ice covered areas.

the experiments represent all possible, and equally plausible, combinations of specified meta-parameters used to quantify model skill of the eight CMIP5 models, based on their ability to simulate ET. Note that some combinations are not possible due to ET dataset temporal coverage, and because regridding using box averaging is used only for upscaling from fine to coarse spatial scales.

For each of the 68 700 benchmarking experiments, we quantify model skill using the root mean squared error (RMSE) and correlation coefficient (ρ) in space and time. These metrics are common in model–data intercomparisons (Blyth *et al* 2011, Cadule *et al* 2010, Schwalm *et al* 2010, Schaefer *et al* 2012, Soares *et al* 2012) although more sophisticated metrics also exist (Braverman *et al* 2011). We also evaluate distributional agreement (S_{time}), the degree of overlap between reference and simulated distributions using discretized probability density functions (Perkins *et al* 2007). This is not as widespread in model evaluation studies but is relevant as the CMIP5 runs evaluated here are initialized several decades before the evaluation period and do not perforce track unforced internal climate variability.

The spatial metrics (ρ_{space} and $\text{RMSE}_{\text{space}}$) are area-weighted and based on the modeled and reference long-term mean by grid cell:

$$\rho_{\text{space}} = \frac{\sum_{i=1}^{n} w_i(y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sqrt{\sum_{i=1}^{n} w_i(y_i - \mu_y)^2}\sqrt{\sum_{i=1}^{n} w_i(\hat{y}_i - \mu_{\hat{y}})^2}} \quad (1)$$

$$\text{RMSE}_{\text{space}} = \sqrt{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2} \quad (2)$$

where y_i and $\hat{y}_i$ are the average observed and simulated values for a grid cell across a given decade (i.e., long-term monthly mean by grid cell), n is the number grid cells, and μ_y and $\mu_{\hat{y}}$ are the spatial means of y_i and $\hat{y}_i$ calculated across n grid cells. Weights are given by w_i; a weighting factor that sums to unity and is based on grid cell area.

The temporal skill metrics (ρ_{time} and $\text{RMSE}_{\text{time}}$) use area-integrated global monthly time series:

$$\rho_{\text{time}} = \frac{\sum_{i=1}^{n}(y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \mu_{\hat{y}})^2}} \quad (3)$$

$$\text{RMSE}_{\text{time}} = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

where y_i and $\hat{y}_i$ are observed and simulated global ET in monthly time series for a given decade, n is the number of months ($n = 120$), and μ_y and $\mu_{\hat{y}}$ are mean values across the full time series. For temporal correlation (equation (3)) we focus on anomalies, with the mean seasonal cycle over the period 1990–1994 removed (time period common to all references/models). For equations (3) and (4) the global values y_i and $\hat{y}_i$ are based on area-integration using w_i as a weighting factor.

Distributional agreement (S_{time}) also uses area-integrated global monthly time series:

$$S_{\text{time}} = \sum_{i=1}^{b} \text{minimum}(Z_{\hat{y},i}, Z_{y,i}) \quad (5)$$

where $Z_{\hat{y},i}$ and $Z_{y,i}$ are the frequency of values in a given bin for simulated (y_i) and reference ($\hat{y}_i$) ET in global monthly anomaly time series, and b is the number of bins. S_{time} is the cumulative minimum value of two distributions across each bin and is a measure of common area between two distributions (Perkins *et al* 2007). Bins are determined using equal spacing across the combined range of simulated and reference values for the target decade. S_{time} values are largely insensitive across a broad range of bin numbers, thus a value of $b = 12$ is used throughout. A value of unity indicates perfect overlap (identical distributions); whereas zero indicates completely disjoint distributions. This is a weaker test than the temporal ρ and RMSE metrics in the sense that an exact temporal matching is not required. S_{time} tracks only if the number of events, e.g., a global monthly anomaly of ET in a given range or bin, that occur over the targeted time period is similar in reference and simulation.

For all metrics both n and w_i are, within a given benchmarking experiment, constant and reference terrestrial vegetated grid cells only. Across benchmarking experiments both n (for spatial metrics only) and w_i change based on which of the six land masks is used. In addition to skill metrics, we also generate model rankings based on inferred skill, i.e., the lowest RMSE and highest ρ or S_{time} values have the 'best' or lowest ranks. By doing so, we are able to investigate the downstream impacts of benchmarking meta-parameter choices on the often-asked question: 'what is the best model?'

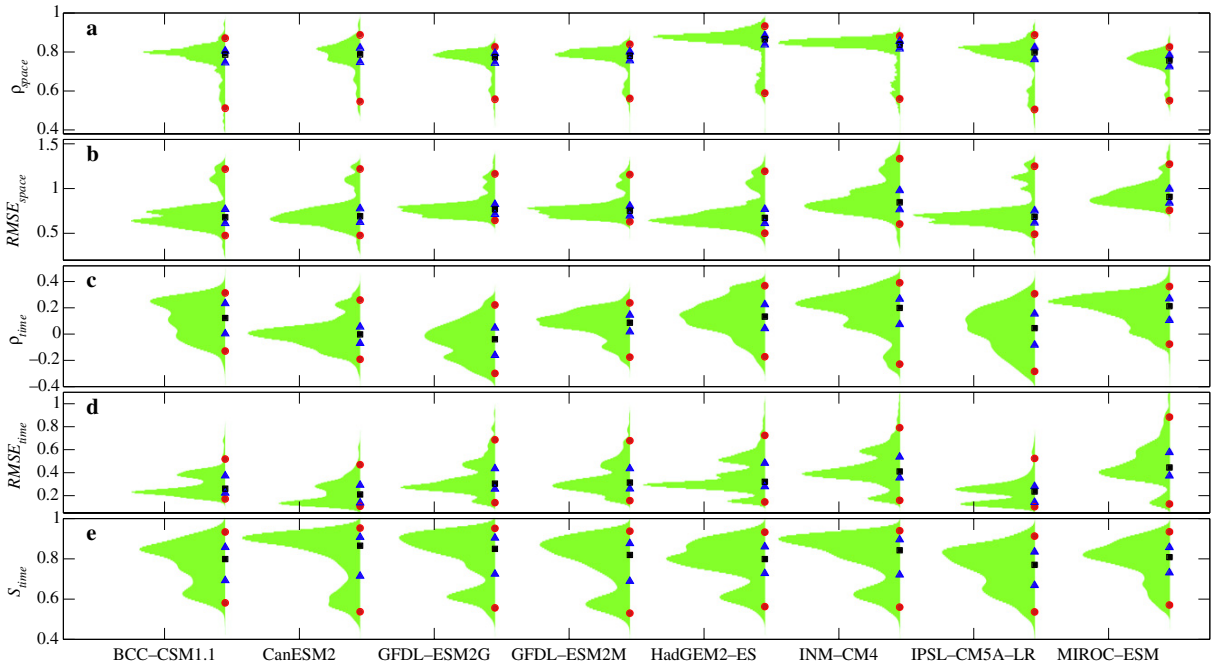

Figure 2. Skill metrics by model. Smoothed histograms for (a) spatial correlation, ρ_{space}; (b) spatial RMSE, $\text{RMSE}_{\text{space}}$; (c) temporal correlation, ρ_{time}; (d) temporal RMSE, $\text{RMSE}_{\text{time}}$; and (e) distributional similarity, S_{time}. Distributions are displayed as probability density functions and share the same scale within each panel. Colored symbols give percentiles. Median, black square; interquartile range (25–75 percentiles), blue triangles; and 2.5–97.5 percentiles, red circles.

Finally, we use all benchmarking experiments for a given model to quantify uncertainty in model skill and rank. Skill metrics, similar to the reference and simulated values, are not fixed and known without error. As uncertainty for these variables is typically not available to be propagated into a skill metric, we derive uncertainty (confidence intervals) in model skill and rank by grouping all skill results by CMIP5 model and extracting relevant percentiles, e.g., a model-specific 95% confidence interval for a given skill metric is derived using the 2.5 and 97.5 percentiles across all benchmarking experiments for that same model.

We quantify the influence of each meta-parameter, as well as the impact of the examined climate model itself, on inferred model skill with a decision tree (Breiman *et al* 1984). These are built by sequentially splitting the data (model skill metrics across all combinations of meta-parameter and climate model in this study) into homogeneous groups. The resulting hierarchy of groups, i.e., the decision tree, is then used to calculate the importance of each meta-parameter and that of the climate models themselves (Breiman *et al* 1984). As the scale for importance is non-intuitive, we derive relative importance by scaling the sum of raw importance scores to 100. Ideally, climate model should have the greatest 'importance', i.e., the greatest impact on inferred model skill, while meta-parameter and climate model choice in the benchmarking experiments should have only a marginal influence on inferred model skill. Such a result would indicate that inferred model rank is robust to the choice of meta-parameters.

3. Results

Inferred model skill varies substantially across the examined climate models, meta-parameters, and metrics (figure 2). Spatial correlation between model and reference product (ρ_{space}) ranges from 0.20 to 0.97 (figure 2(a)). The spatially-weighted RMSE ($\text{RMSE}_{\text{space}}$) varies from 0.25 to 1.5 mm d^{-1} (figure 2(b)); a wide range given the spread in reference ET fluxes (supplementary table 1) from 1.3 to 1.8 mm d^{-1}. Temporal correlation (ρ_{time}) ranges from -0.36 to $+0.53$ (figure 1(c)), i.e., for some sets of meta-parameters reference and simulation are anti-correlated. $\text{RMSE}_{\text{time}}$ (figure 2(d)), which is generally less than $\text{RMSE}_{\text{space}}$, varies between 0.08 and 1.0 mm d^{-1} or 5 and 65% of the mean reference value. Distributional agreement (S_{time}) for monthly anomalies shows uniformly higher levels of model skill (figure 2(e)) than their correlation (ρ_{time}). This is expected as S_{time} is a weaker test, i.e., high skill levels require only congruence in the number of occurrences in a given range or distributional bin as opposed to the exact temporal sequencing needed for ρ_{time}. While these large observed ranges in model skill suggest multiple skill levels for a given model, it is noteworthy that these ranges are solely attributable to how the intercomparison is performed.

Using clusters of grid cells (e.g., geographic region, plant functional types, climatic zones) to control for land surface heterogeneity does not lessen the range in inferred model skill (e.g., ρ_{space}; supplementary figure 1 available at stacks.iop.org/ERL/8/024028/mmedia) and we therefore limit our discussion to global results. Similarly, although the decadal

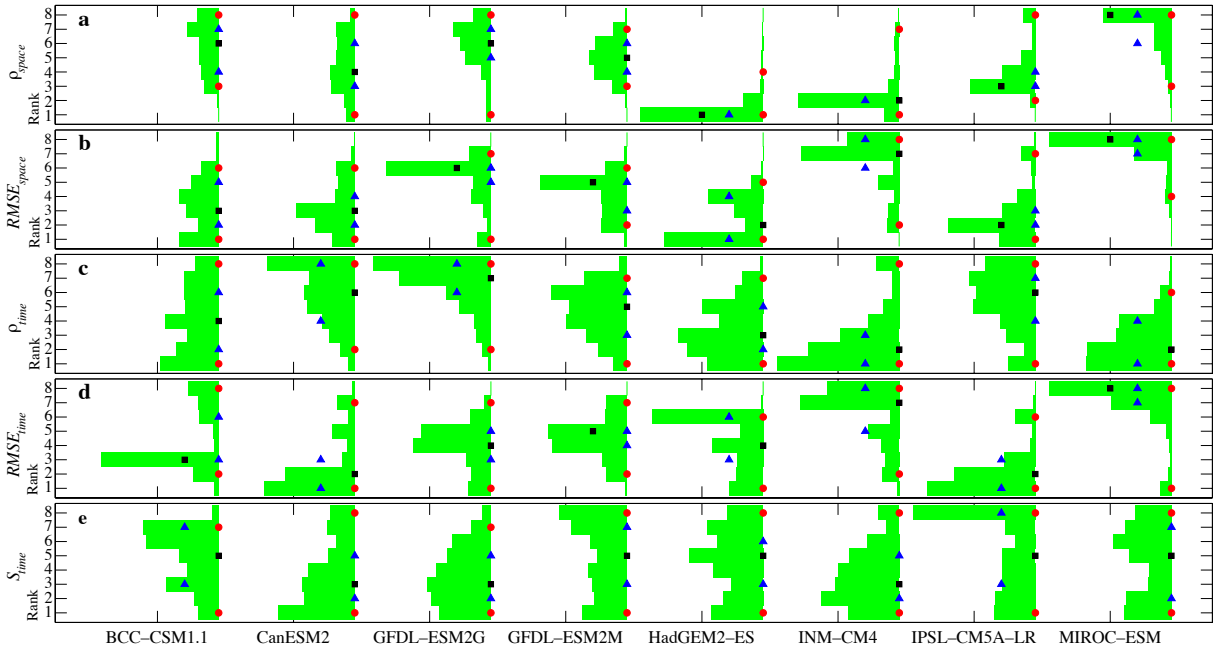

Figure 3. Skill rank by model. Histograms for ranked (a) spatial correlation, ρ_{space}; (b) spatial RMSE, $\text{RMSE}_{\text{space}}$; (c) temporal correlation, ρ_{time}; (d) temporal RMSE, $\text{RMSE}_{\text{time}}$; and (e) distributional similarity, S_{time}. Lower ranks denote relatively higher levels of model–data agreement. Distributions are displayed as horizontal histograms and share the same scale within each panel. Colored symbols give percentiles. Median, black square; interquartile range (25–75 percentiles), blue triangles; and 2.5–97.5 percentiles, red circles. Some symbols jittered to avoid overlap.

time periods overlap, suggesting a loss in degrees of freedom in estimating confidence bounds, we find the distributions for overlapping and non-overlapping decades highly similar (supplementary figure 2 available at stacks.iop.org/ERL/8/024028/mmedia). As only four of the six ET references extend to multiple (i.e., two) non-overlapping decades, the use of overlapping decades allows for a ten-fold increase in benchmarking experiments. We therefore retain all possible overlapping decades in our discussion.

To identify plausible bounds of model skill, 95% confidence intervals (2.5 and 97.5 percentiles) and the interquartile range (25 and 75 percentiles) for inferred model skill are derived assuming all sets of meta-parameters are equally valid (figure 2). The 95% confidence intervals overlap across all climate models for each of the five examined metrics, precluding clear ranking of the models. In some cases, the model with the 'best' 95% confidence interval upper limit (high ρ and S_{time} or low RMSE) is not the same as the model with the 'best' interquartile range upper limit (e.g., INM-CM4 and MIROC-ESM for $\text{RMSE}_{\text{space}}$ (figure 2(b))). As a result, a clear determination of ranking in model skill is not possible. Even though the 95% confidence intervals are obviously narrower than the full range of inferred skill, these ranges are too wide to address model skill. This ambiguity is problematic for benchmarking, where the ultimate aim is to diagnose shortcomings in model characteristics. A model simultaneously showing high and low levels of agreement across equally plausible benchmarking meta-parameter choices hampers any efforts at diagnosing model deficiencies.

Consistent with the inferred model skill results, the inferred rank of individual models also varies dramatically across meta-parameter choices (figure 3), precluding the assignment of a single rank to any model. For 35 of the 40 climate model × metric combinations, all ranks are observed. Nevertheless, some models generally do better (rank distribution mode of 1, e.g., IMN-CM4 for ρ_{time} rank (figure 3(c)) and Can-ESM2 for S_{time} (figure 3(e))) or worse (mode of 8, e.g., MIROC-ESM for ρ_{space} and $\text{RMSE}_{\text{space}}$ ranks (figures 3(a) and (b) respectively)) for some metrics. Such tendencies are however not consistent for a given model across all metrics (e.g., IPSL-CM5A-LR for $\text{RMSE}_{\text{space}}$ versus S_{time} ranks (figures 3(b) and (e) respectively)). This implies that although qualitative comparisons between models for specific metrics may be possible in some cases, model rank is best represented by a discrete probability mass function rather than by a scalar value.

As with the raw metric values, we use the 95% confidence intervals and interquartile range to identify plausible bounds on model rank. Across the 40 combinations of metrics and climate models, all but three combinations span ranks 3 through 6 at the 95% confidence level, and all but ten combinations span ranks 2 through 7. The interquartile ranges for model rank are substantially narrower, however, ranging from a single plausible rank (e.g., HadGEM2-ES and INM-CM4 for ρ_{space}) to five plausible ranks (e.g., BCC-CSM1.1 and Can-ESM2 for ρ_{time}).

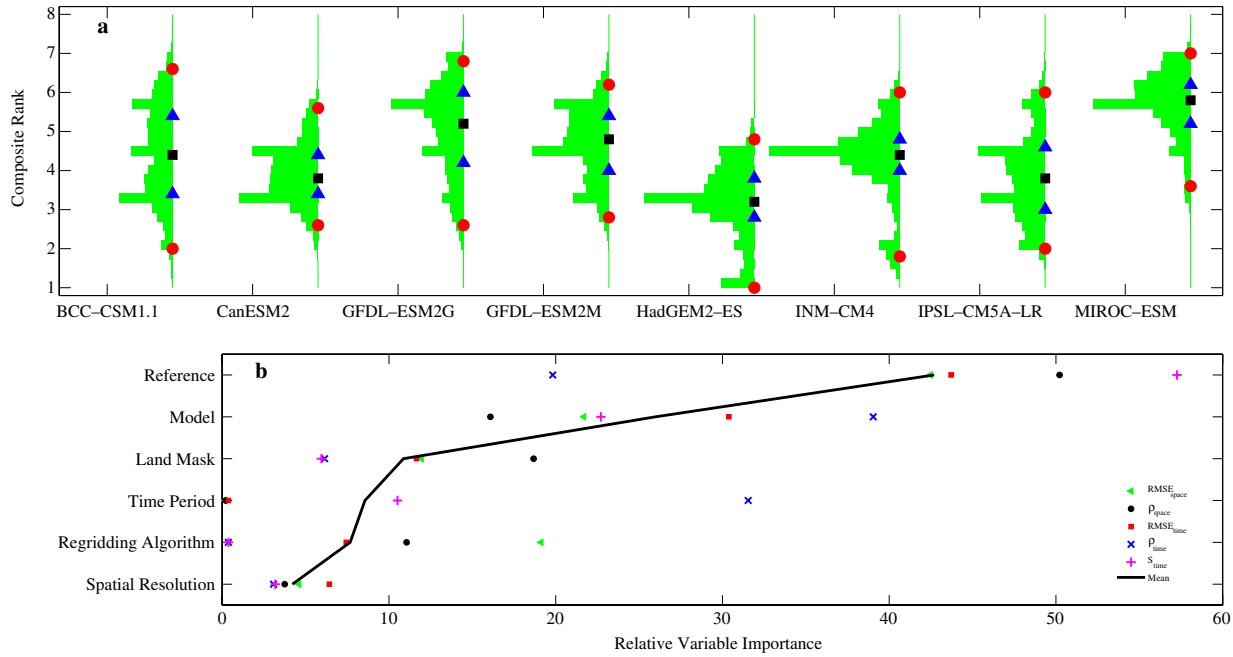

Figure 4. Composite rank and variable importance. (a) Mean rank across all ranked skill metrics. All values, in 0.2 step increments from 1 to 8, shown. Colored symbols give percentiles. Median, black square; interquartile range (25–75 percentiles), blue triangles; and 2.5–97.5 percentiles, red circles. (b) Relative variable importance by skill metric and on average for each meta-parameter and model.

Averaging ranks across all five metrics (figure 4(a)) provides a more complete view of model skill. This type of composite metric generalizes to multiple variables with variable weights. For this case study we use a composite rank based on equal weighting. This generally yields more symmetric distributions, but even the interquartile ranges on rank do not converge on a single inferred overall rank for any model. This suggests that both the basic question 'what is the best model?' and the more specific question 'how much confidence can be placed in model simulations?' do not have clear answers given the observed uncertainty in inferred model skill.

Despite the lack of a single representative model rank, some models are more likely to perform better than others. For example, HadGEM2-ES is the only model with a 95% confidence interval that includes an aggregated rank of one (figure 4(a)). Other models (e.g., MIROC-ESM) have both a high probability of a poor ranking, and a low probability of a good ranking. Such probabilistic information allows for a fuller characterization of model skill and can only be obtained through a factorial approach to benchmarking as applied here.

The decision tree analysis (figure 4(b)) shows that the choice of reference dataset is the most important factor in determining inferred model skill. This is primarily because differences in reference datasets (range: 60–85 10^3 km^3 yr^{-1}) are large relative to differences in climate model estimates (range: 66–87 10^3 km^3 yr^{-1}). This holds for all metrics except ρ_{time} (figure 4(b)), where model and time period choice are more important than reference dataset. Second in overall importance, and considerably more important than the remaining meta-parameters, is the choice of model. This applies to all metrics except ρ_{space} (figure 4(b)) where land mask ranks only behind reference dataset in importance.

Although reference dataset is the key determinant for model skill distributions, the overall variability in model skill is not attributable to a specific reference product itself. We show this by holding both CMIP5 model and reference product constant for model skill (figure 5) and rank (figure 6). Generally there is a single reference product that alone spans the full range, or nearly so. This is more pronounced for spatial skill metrics (figure 5) and ranks (figure 6). For temporal skill metrics and S_{time} this feature is less prominent but even here there is substantial overlap in skill distribution. In no case are any distributions completely disjoint; S_{time} for CAN-ESM2, GFDL-ESM2G, and GFDL-EMS2M and ρ_{time} for INM-CM4 have the lowest distributional overlap, i.e., nearly disjoint distributions (figure 5). Also, where a one-number summary of skill, i.e., the median value, would indicate a gradient in skill attributable to reference (e.g., HadGEM2-ES for RMSE$_{\text{time}}$ (figure 5) or GFDL-ESM2G for S_{time} rank (figure 6)) the full distributions show extensive overlap in skill and rank. Overall, even though reference is the largest mode of model skill variability, other meta-parameters are associated with significant variation in skill.

4. Conclusion

Confronting models with observationally-based references as a means to assess model skill is an integral part of model development. Here we show that, across multiple sets of plausible benchmarking meta-parameters, that inferred model skill and rank are highly variable and uncertain.

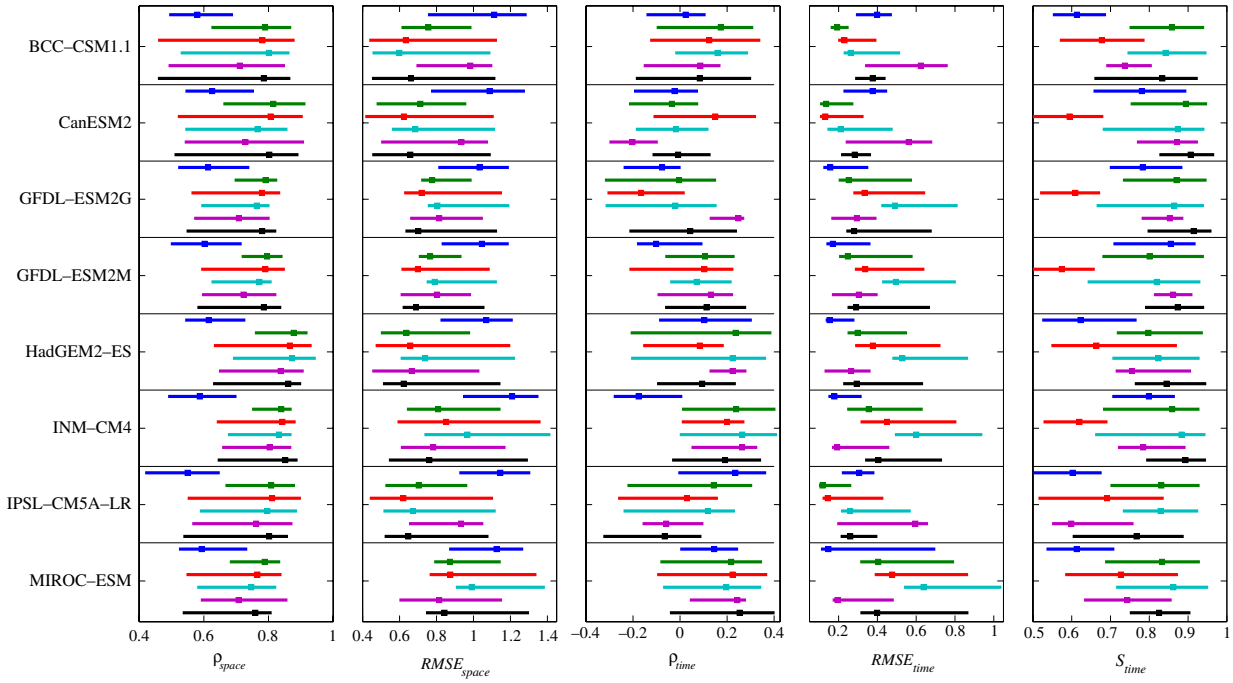

Figure 5. Range in model skill by CMIP5 model and ET reference. Columns show compact horizontal boxplots for a given model skill metric. Median, square; and 2.5–97.5 percentiles, thick line. Colors denote ET reference product: blue, AWB; green, CSIRO; red, MPI; cyan, NTSG; magenta, PT-JPL; and black, UDEL. Rows show each CMIP5 model.

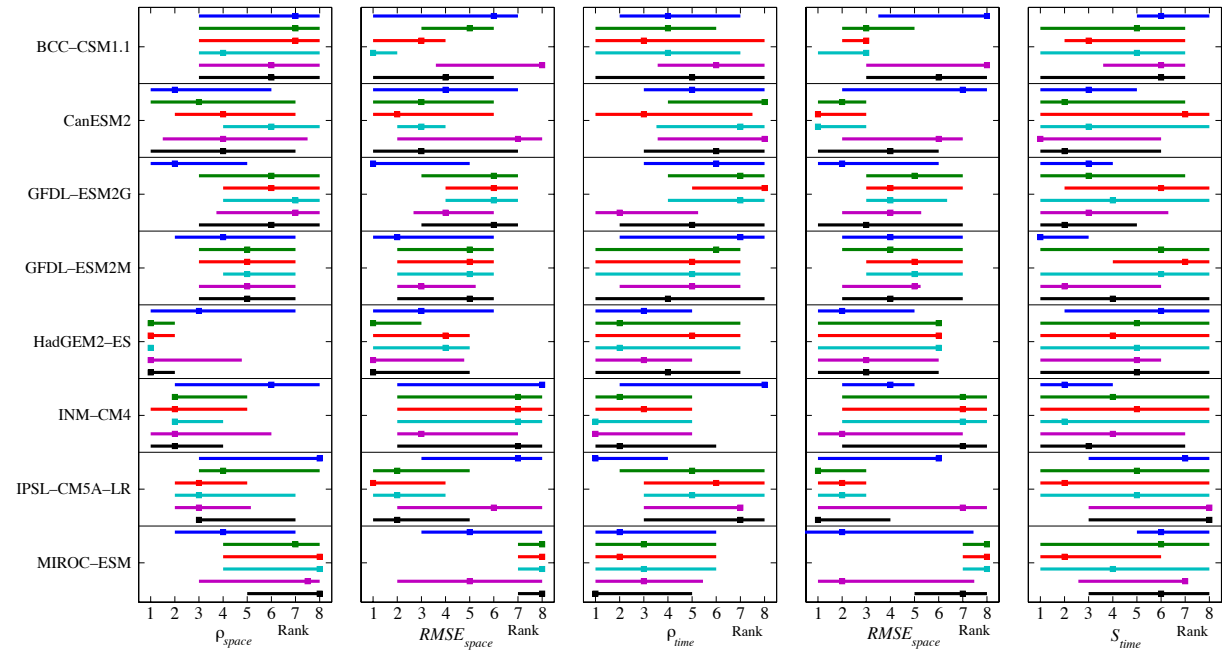

Figure 6. Range in model rank by CMIP5 model and ET reference. Columns show compact horizontal boxplots for a given model skill metric. Median, square; and 2.5–97.5 percentiles, thick line. Colors denote ET reference product: blue, AWB; green, CSIRO; red, MPI; cyan, NTSG; magenta, PT-JPL; and black, UDEL. Rows show each CMIP5 model.

This is problematic in a benchmarking context as a model simultaneously showing multiple levels of model skill/rank across equally plausible meta-parameters precludes a diagnosis of model deficiencies. For this case study, the main driver of uncertainty in model skill is the reference ET dataset chosen for the evaluation.

This study does not include estimates of uncertainty from the models or the reference data products, as these

estimates are not universally available. However, doing so would broaden the range of plausible model skill or model rank for any given chosen reference. As a result, this study represents a conservative assessment of our ability to rank models based on their skill level relative to a single reference data product or a suite of reference data.

A key implication from this study for future model intercomparison projects and community benchmarking efforts, such as ILAMB (International Land Model Benchmarking project; http://ilamb.org/) and the WGNE/WGCM (Working Group on Numerical Experimentation and Working Group on Coupled Modeling, respectively) Climate Model Metrics Panel (www-metrics-panel.llnl.gov/wiki), is that the choice of reference dataset could potentially have more influence on inferred model skill or rank than the model being evaluated. Furthermore, our results strongly suggest that model skill is partially decoupled from intrinsic model characteristics. While the benchmarking experiments here focus solely on ET, we expect similar ambiguity for other biogeochemical and biophysical variables where multiple reference products are available. This indicates that substantial time and effort must be spent in developing community-accepted standard reference datasets with emphasis on quality control and robust uncertainty quantification (e.g., GEWEX LandFlux/LandFlux-EVAL (Mueller *et al* 2011)). More generally, evaluating the reference datasets themselves is a critical step towards decreasing the ambiguity in inferred model skill and/or ranks.

Finally, given the large variability in inferred model skill/rank, one-number summaries of model–data mismatch may be misleading and erroneous. Instead, model rank and skill should be presented probabilistically rather than as single summary values. Although point estimates of skill or rank may have value in characterizing the central tendency of model skill, because of the sensitivity of inferred skill/rank to benchmarking choices, it is inadvisable to rely solely on such scores to inform model development.

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 US 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. CRS, DNH, and AMM were supported by the National Aeronautics and Space Administration (NASA) under Grant No. NNX10AG01A 'The NACP *Multi-Scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP)*'. CRS was also supported by NASA Grant No. NNX12AK12G. JBF contributed to this paper at the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA.

References

Abramowitz G 2005 Towards a benchmark for land surface models *Geophys. Res. Lett.* **32** L22702

Abramowitz G 2012 Towards a public, standardized, diagnostic benchmarking system for land surface models *Geosci. Model Dev.* **5** 819–27

Blyth E, Clark D B, Ellis R, Huntingford C, Los S, Pryor M, Best M and Sitch S 2011 A comprehensive set of benchmark tests for a land surface model of simultaneous fluxes of water and carbon at both the global and seasonal scale *Geosci. Model Dev.* **4** 255–69

Braverman A, Cressie N and Teixeira J 2011 A likelihood-based comparison of temporal models for physical processes *Stat. Anal. Data Min.* **4** 247–58

Breiman L, Friedman J, Olshen R and Stone C 1984 *Classification and Regression Trees* (Boca Raton, FL: CRC Press)

Cadule P, Friedlingstein P, Bopp L, Sitch S, Jones C D, Ciais P, Piao S L and Peylin P 2010 Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements *Glob. Biogeochem. Cycles* **24** Gb2016

Fisher J B, Tu K P and Baldocchi D D 2008 Global estimates of the land-atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites *Remote Sens. Environ.* **112** 901–19

Friedlingstein P *et al* 2006 Climate-carbon cycle feedback analysis: results from the (CMIP)-M-4 model intercomparison *J. Clim.* **19** 3337–53

Gleckler P J, Taylor K E and Doutriaux C 2008 Performance metrics for climate models *J. Geophys. Res.* **113** D06104

Jiménez C *et al* 2011 Global intercomparison of 12 land surface heat flux estimates *J. Geophys. Res.* **116** D02102

Jung M, Henkel K, Herold M and Churkina G 2006 Exploiting synergies of global land cover products for carbon cycle modeling *Remote Sens. Environ.* **101** 534–53

Jung M *et al* 2011 Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations *J. Geophys. Res.* **116** G00J07

Loveland T R, Reed B C, Brown J F, Ohlen D O, Zhu J, Yang L and Merchant J W 2001 Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data *Int. J. Remote Sens.* **21** 1303–30

Luo Y Q *et al* 2012 A framework for benchmarking land models *Biogeosciences* **9** 3857–74

Meehl G A *et al* 2007 The WCRP CMIP3 multi-model dataset: a new era in climate change research *Bull. Am. Meteorol. Soc.* **88** 1383–94

Mueller B *et al* 2011 Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations *Geophys. Res. Lett.* **38** L06402

Perkins S E, Pitman A J, Holbrook N J and McAneney J 2007 Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature and precipitation over Australia using probability density functions *J. Clim.* **20** 4356–76

Randall D A *et al* 2007 Climate models and their evaluation *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* ed S Solomon, D Qin, M Manning, Z Chen, M Marquis, K B Averyt, M Tignor and H L Miller (Cambridge: Cambridge University Press)

Randerson J T *et al* 2009 Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models *Glob. Change Biol.* **15** 2462–84

Schaefer K *et al* 2012 A model-data comparison of gross primary productivity: results from the North American carbon program site synthesis *J. Geophys. Res.* **117** G03010

Schwalm C R, Williams C A and Schaefer K M 2011 Carbon consequences of global hydrologic change, 1948–2009 *J. Geophys. Res.* **116** G03042

Schwalm C R *et al* 2010 A model-data intercomparison of CO_2 exchange across North America: results from the North American carbon program site synthesis *J. Geophys. Res.* **115** G00H05

Soares P M M, Cardoso R M, Miranda P M A, Viterbo P and Belo-Pereira M 2012 Assessment of the ENSEMBLES regional climate models in the representation of precipitation variability and extremes over Portugal *J. Geophys. Res.* **117** D07114

Taylor K E, Stouffer R J and Meehl G A 2012 An overview of CMIP5 and the experiment design *Bull. Am. Meteorol. Soc.* **93** 485–98

Vinukollu R K, Wood E F, Ferguson C R and Fisher J B 2011 Global estimates of evapotranspiration for climate studies using multi-sensor remote sensing data: evapotranspiration–remote sensing and modeling evaluation of three process-based approaches *Remote Sens. Environ.* **115** 801–23