water vapor was produced in high purity in the case of full CH_4 conversion. Because a sweep gas was used, the CO in the product stream was diluted, and a molar fraction of 0.56 rather than 0.85 was obtained. However, this is still 30% higher than the equilibrium molar fraction of CO in conventional dry reforming at the same conditions (Fig. 4A). Moreover, our goal was not to optimize the flow of sweep gas, but rather to highlight the possibilities based on a combination of known processes. Hence, it may be possible to further increase the yield and purity of CO by fine-tuning process variables, reactor configuration, or materials. This flexibility is not evident at all for the conventional DRM process, which is limited in yield by the WGS reaction (Eq. 2).

Exergy calculations (which account for irreversible losses) were performed to estimate the energetics of the proposed process (Fig. 4B and fig. S23B) (*20*). Along with the advantage of improved product purity, super-dry reforming of CH_4 also resulted in a very low exergy destruction per mole CO_2 converted (Fig. 4B). Indeed, the exergy destruction for CO_2 conversion is up to 25 to 50% lower as compared with that of conventional DRM. The latter typically requires operating temperatures of 1073 to 1273 K in order to reach high-equilibrium conversion of CH_4 and CO_2 while minimizing the thermodynamic driving force for carbon formation (*13*). Compared with these requirements, super-dry reforming shows both practical and economic benefits. For one, the requirement of a noble metal catalyst in order to mitigate excessive carbon formation (*13*) becomes obsolete because conditions of super-dry reforming are thermodynamically unfavorable for carbon deposition. Hence, the application of cheaper Ni-based catalysts is perfectly feasible, even at temperatures as low as 1023 K.

The product streams of super-dry reforming have a higher purity than those of conventional CH_4 reformers, which should reduce separation costs. Moreover, by applying Le Chatelier's principle to lift CO_2 utilization efficiency, each molecule of CH_4 theoretically allows the conversion of three molecules of CO_2, making super-dry reforming of CH_4 an intensified process for isothermal CO_2 utilization. The presence of a CO_2 acceptor, in this case CaO, allows for the application of renewable feedstocks such as biogas as a reducing agent for iron oxide reduction.

REFERENCES AND NOTES

1. L. Li, N. Zhao, W. Wei, Y. Sun, *Fuel* **108**, 112–130 (2013).
2. C. Song, *Catal. Today* **115**, 2–32 (2006).
3. M. Aresta, A. Dibenedetto, A. Angelini, *J. CO_2 Util.* **34**, 65–73 (2013).
4. M. E. Boot-Handford *et al.*, *Energ. Environ. Sci.* **7**, 130–189 (2014).
5. V. V. Galvita, H. Poelman, G. B. Marin, *J. Power Sources* **286**, 362–370 (2015).
6. J. C. Liao, L. Mi, S. Pontrelli, S. Luo, *Nat. Rev. Microbiol.* **14**, 288–304 (2016).
7. S. Lin *et al.*, *Science* **349**, 1208–1213 (2015).
8. M. Schreier *et al.*, *Nat. Commun.* 10.1038/ncomms8326 (2015).
9. J. R. Scheffe, J. Li, A. W. Weimer, *Int. J. Hydrogen Energy* **35**, 3333–3340 (2010).
10. J. R. Scheffe, A. Steinfeld, *Mater. Today* **17**, 341–348 (2014).
11. C. L. Muhich *et al.*, *Science* **341**, 540–542 (2013).
12. T. Kodama, N. Gokon, *Chem. Rev.* **107**, 4048–4077 (2007).
13. D. Pakhare, J. Spivey, *Chem. Soc. Rev.* **43**, 7813–7837 (2014).
14. J. Hu *et al.*, *J. CO_2 Util.* **16**, 8–16 (2016).
15. V. V. Galvita, H. Poelman, C. Detavernier, G. B. Marin, *Appl. Catal. B* **164**, 184–191 (2015).
16. O. Muraza, A. Galadima, *Int. J. Energy Res.* **39**, 1196–1216 (2015).
17. N. V. R. A. Dharanipragada *et al.*, *J. Mater. Chem. A* **3**, 16251–16262 (2015).
18. S. A. Theofanidis, V. V. Galvita, H. Poelman, G. B. Marin, *ACS Catal.* **5**, 3028–3039 (2015).
19. D. P. Harrison, *Ind. Eng. Chem. Res.* **47**, 6486–6501 (2008).
20. Materials and methods are available as supplementary materials on *Science* Online.
21. F. Jiao *et al.*, *Science* **351**, 1065–1068 (2016).
22. H. M. Torres Galvis *et al.*, *Science* **335**, 835–838 (2012).
23. P. Heidebrecht, K. Sundmacher, *Chem. Eng. Sci.* **64**, 5057–5065 (2009).

ACKNOWLEDGMENTS

This work was supported by the Long Term Structural Methusalem Funding of the Flemish Government, the Interuniversity Attraction Poles Programme, IAP7/5, Belgian State–Belgian Science Policy, and the Fund for Scientific Research Flanders (FWO; project G004613N). L.C.B. acknowledges financial support from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT Vlaanderen). Additionally, he thanks his colleagues A. Dharanipragada and S. Theofanidis for supplying the oxygen carrier material and CH_4 reforming catalyst. We thank V. Bliznuk (Department of Materials Science and Engineering of Ghent University) for his technical support concerning STEM analysis, G. Rampelberg for help with in situ XRD (Department of Solid State Sciences, Ghent University), and O. Janssens (Department of Solid State Sciences, Ghent University) for performing ex situ XRD and SEM measurements. All data are presented in the main text and supplementary materials.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/354/6311/449/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S23
Tables S1 to S4
References (*24–39*)

4 August 2016; accepted 26 September 2016
10.1126/science.aah7161

CLIMATE DATA

Using climate models to estimate the quality of global observational data sets

François Massonnet,[1,2]* Omar Bellprat,[1] Virginie Guemas,[1,3] Francisco J. Doblas-Reyes[1,4]

Observational estimates of the climate system are essential to monitoring and understanding ongoing climate change and to assessing the quality of climate models used to produce near- and long-term climate information. This study poses the dual and unconventional question: Can climate models be used to assess the quality of observational references? We show that this question not only rests on solid theoretical grounds but also offers insightful applications in practice. By comparing four observational products of sea surface temperature with a large multimodel climate forecast ensemble, we find compelling evidence that models systematically score better against the most recent, advanced, but also most independent product. These results call for generalized procedures of model-observation comparison and provide guidance for a more objective observational data set selection.

There is now overwhelming evidence that Earth's climate has changed at an unusually rapid pace during the last century, that these changes bear a clear human signature, and that they will be enhanced if anthropogenic emissions continue unabated (*1*). The development of large-scale observational networks has been a major advance to reaching such levels of evidence. Observations of essential climate variables [e.g., sea surface temperature (SST), sea ice extent (*2*)] are indeed central for the study of climate variability (*1*), for detection and attribution of human-induced climate change (*1*, *3*), and for constraining long-term projections (*1*, *4*). Major international and coordinated observing programs are currently underway to continue these efforts (*5*). However, with the emergence of multiple observational references (ORs), sometimes divergent, a natural question arises: What is the underlying quality of these products? A direct answer to this question is not easily achieved because there is by definition no universal knowledge of the true state of our climate (*6*).

Here we present a framework for the evaluation of ORs addressing this gap. The approach relies on the use of climate models taken as references, and not as subjects of assessment as

[1]Earth Sciences Department, Barcelona Supercomputing Center–Centro Nacional de Supercomputación (BSC-CNS), Barcelona, Spain. [2]Georges Lemaître Centre for Earth and Climate Research (TECLIM), Earth and Life Institute (ELI), Université catholique de Louvain, Louvain-la-Neuve, Belgium. [3]Centre National de Recherches Météorologiques (CNRM), Toulouse, France. [4]ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain.
*Corresponding author. Email: francois.massonnet@bsc.es

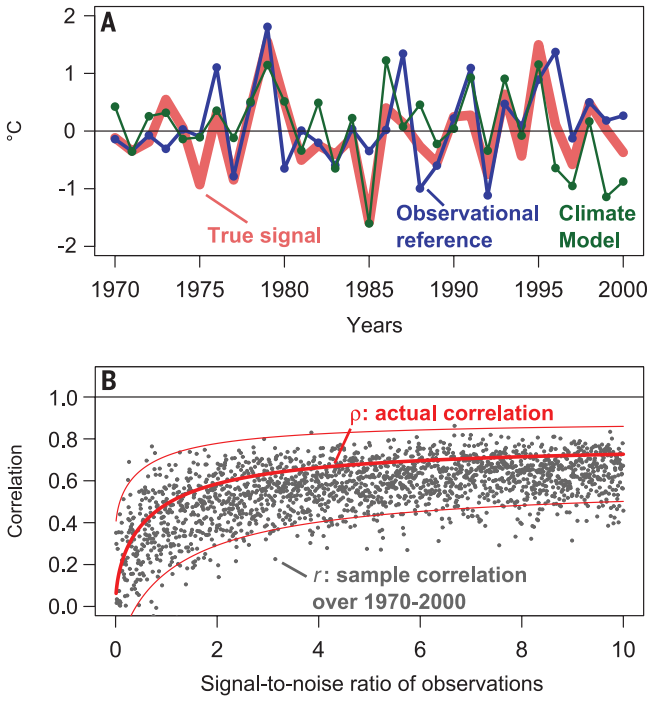

Fig. 1. A simple statistical toy model (Eq. 1) for understanding the dependence of correlation on observational error. (**A**) A true but unknown signal ϵ (light red) mimicking, for example, annual mean SST anomalies over some region is generated over 1970 to 2000 as a white-noise process with SD σ_ϵ = 0.8°C. An observationally based reference (blue) systematically underestimating the true variability (α_{OR} = 0.8) is then sampled according to Eq. 1 with SD σ_{OR} = 0.6°C. Finally, a climate model estimate (green) is shown with multiplicative bias α_{CM} = 0.7 and with its own and irreducible errors having SDs $\sigma_{CM,p}$ = 0.2°C and $\sigma_{CM,i}$ = 0.4°C, respectively. (**B**) Dependence of the linear correlation coefficient on the signal-to-noise ratio of observations $\gamma_{OR} = \sigma_\epsilon^2/\sigma_{OR}^2$. The red lines show the analytical, predicted dependence (Eq. 2) and the 95% confidence interval using Fisher's *z* transform (*32*). Gray dots are sample correlations between generated observation and model data for different levels of observational error, keeping the true signal unchanged [as in (A)] but each time generating a new model realization (with fixed error statistics α_{CM} = 0.7, $\sigma_{CM,p}$ = 0.2°C and $\sigma_{CM,i}$ = 0.4°C) and an observation (α_{OR} = 0.8) while varying the SE so that a desired signal-to-noise ratio is achieved.

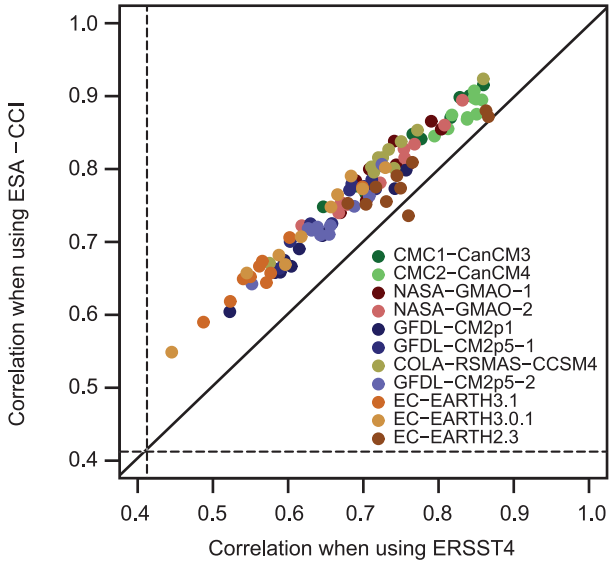

Fig. 2. Systematic dependence of correlation on verification product. Each dot highlights the relationship between the correlation of one August SST forecast verified against the ESA-CCI product (*y* axis; product based on satellite information only) and the correlation of the same forecast verified against the ERSST4 product (*x* axis; product based on in situ information only). Reference period is 1993 to 2009. There are 110 forecasts in total (11 models, 10 members each). The solid line is the 1:1 line and delimits regions where ESA-CCI or ERSST4 scores better. The dashed lines are the threshold above which correlations are significant at the 0.05 level (the 1-year autocorrelations of the time series considered are weak and not significant).

has been widely done in the past (*7*, *8*). The rationale behind this approach relies on the so-called "truth-plus-noise" paradigm (*9*–*14*), which assumes that observations and models are both noisy versions of the true (but unknown) state of the climate system. In that view, observations and models play symmetrical roles so that it is possible to use one to estimate how close the other is from the true state, and vice versa. In line with this paradigm, we claim that climate models can be appropriate tools for estimating the quality of ORs. We accumulate the necessary evidence in three steps. First, we rely on elementary logic and take advantage of the symmetry of common metrics of model performance. Then, we show with a simple statistical toy model how observational error can degrade model performance (symmetrically to model error), turning this into an opportunity to reverse the process of model evaluation into one of OR evaluation. Finally, we apply the proposed procedure to a realistic test case involving simulations conducted with large-scale general circulation models and a set of ORs.

"Quality" is here measured in terms of a primary metric of performance widely used in the climate community: the (Pearson) linear correlation between two variables (*15*). Because this metric of performance is mathematically symmetric, the process of evaluation is also inherently symmetric. Provided that this metric of performance is appropriate to characterizing the quality of climate models or forecast systems (*16*–*18*), it also becomes a way to measure the quality of ORs. The same argument holds for other metrics, provided that they respect the underlying hypothesis of symmetry.

To frame this idea in a more concrete context, consider the following simple, yet generic enough, toy model (*19*) consisting of an observational reference (X_{OR}) and model-based (X_{CM}) estimate of some true, but unknown, climate variable ϵ:

$$\begin{cases} X_{OR} = \alpha_{OR}\epsilon + \eta_{OR} \\ X_{CM} = \alpha_{CM}\epsilon + \eta_{CM,p} + \eta_{CM,i} \end{cases} \quad (1)$$

The observational reference X_{OR} is related to the true signal by a multiplicative bias factor $\alpha_{OR} > 0$, that accounts, e.g., for the systematic errors in retrieval from the raw measurement to the final product. The OR estimate is further polluted by a random error η_{OR} that reflects instrumental and sampling errors. Another estimate X_{CM} of the true variations can be obtained through a climate model (see Fig. 1A for an example). Like the OR, the climate model does not necessarily capture the correct amplitude of the variability (hence a multiplicative bias term $\alpha_{CM} > 0$ to model, e.g., the possible incorrect response to external forcings) and is subject to random error $\eta_{CM,p}$ (because of, e.g., unresolved or misrepresented processes including forcings) and random error $\eta_{CM,i}$ due to the inherent chaotic nature of climate dynamics and its intrinsic unpredictability (*20*). All error terms are assumed to follow Gaussian distributions with zero means and known standard deviations. Focusing on the evaluation of anomalies is motivated by the fact that constant offsets between ORs and climate model time series are usually removed by classical linear bias correction techniques. In that sense, we focus on the ability of ORs and climate models to reproduce variability rather than mean states. Finally, all error terms are assumed to be uncorrelated with each other. Under these ideal

conditions, the actual correlation between X_{OR} and X_{CM} follows the expression

$$\rho = \frac{1}{\sqrt{\left(1+\frac{1}{\alpha_{OR}^2\gamma_{OR}}\right)\cdot\left(1+\frac{1}{\alpha_{CM}^2\gamma_{CM}}\right)}} \quad (2)$$

where γ_{OR} and γ_{CM} denote the signal-to-noise ratios of the observation and the model, respectively (see the supplementary text for the demonstration of the relationship and Fig. 1B caption for the definition of signal-to-noise ratios). The relationship, displayed graphically in Fig. 1B for a given model error level, highlights that observational and model error play exchangeable roles in the definition of correlation: Correlation will decrease whenever either observational or model error increases. Despite its simplicity, the toy model illustrates two essential points: (i) Models and ORs are two sides of the same coin and make the process of evaluation intimately intertwined; and (ii) model error is not the only cause of poor model performance: Observational error also contributes to poor correlations. We now investigate this aspect in a more realistic case.

We compiled data from 11 state-of-the-art climate models for which 10-member-ensemble May to August (4-month-long) retrospective predictions of average SST in the Niño3.4 box (120°W to 170°W, 5°S to 5°N) were available between 1993 and 2009 (see supplementary materials for the choice of the period and a thorough description of this ensemble). Unlike the toy model presented above, forecast errors are not independent from each other (fig. S2). Besides, we computed monthly-mean SSTs in the Niño3.4 box for four ORs at various resolutions and based on various methods of retrieval (see supplementary materials for a complete description): ESA-CCI (~0.05° resolution, satellite data only), ERA-Interim (~0.7°, an atmospheric reanalysis also providing SST information), HadISST (1°, based on in situ data but interpolated with satellite data), and ERSST4 (2°, in situ data only). We find that the choice of OR used to verify the forecasts has a systematic and nonnegligible influence on the correlation between the forecast and the OR (Fig. 2). All August forecasts but one score better when assessed against the ESA-CCI product instead of ERSST4, and differences in correlation are commensurate with changes between model versions. For the example given in Fig. 2, the average change in Niño3.4 SST correlation when using ESA-CCI instead of ERSST4 is +0.07 (minimum = −0.02; maximum = 0.11). As a comparison, the mean change in correlations between EC-Earth 3.0 and EC-Earth 3.1 (two model versions) is 0.05 when assessed against ESA-CCI.

To obtain a broader picture, we recorded for each OR the number of forecasts amongst the 440 available (11 models, 10 members each, 4 months of forecasts) that reached the highest correlation with that particular OR (Fig. 3). The results show that forecasts, as a group, tend to favor one particular OR (ESA-CCI), which also happens to be the most recent, at the highest resolution and resorting to the most advanced technology for SST retrieval (*21*). In addition, none of the forecast systems uses the ESA-CCI SST OR for initialization, postprocessing, or validation, which excludes the possibility of reaching high correlation only by construction (*22*). At the other end of the diagram lies ERSST4, the OR that is arguably least adapted for this exercise of SST comparison in the middle of the Pacific Ocean, where sampling is poorer than in other regions of the Pacific (*23*). It is also worth noting that ORs that are a priori not relevant for SST evaluation (ERA-Interim, an atmospheric reanalysis) can score high for the wrong reasons. Five of the 11 models considered are initialized with the same SST fields as those used in ERA-Interim, making the evaluation dependent through a methodological artefact. This OR was deliberately included in our analysis to show that high correlations are a necessary, but not sufficient, indicator of high observational quality.

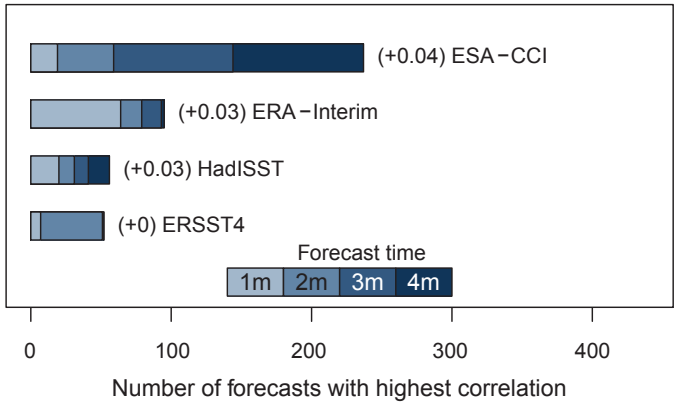

Fig. 3. Impact of the choice of OR on the assessment of forecast quality. A total of 440 seasonal retrospective forecasts of SST in the Niño3.4 box (11 models, 10 members each integrated for 4 months starting from 1 May) are correlated with four ORs of SST for verification over 1993 to 2009. For each OR, we record the number of times that this OR yields the highest correlation. Numbers in parentheses before each OR indicate the average increase in correlation when using that OR instead of ERSST4. Corresponding figures for an alternative metric (RMSE) and an alternative test case (sea ice) can be found in the supplementary materials.

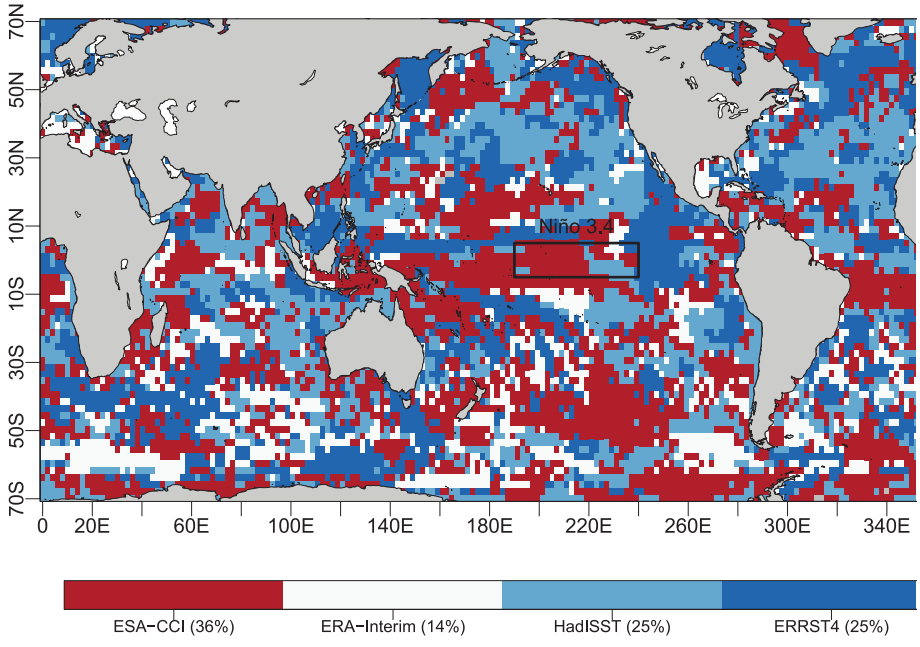

Fig. 4. Spatial distribution of observational reference quality. A total of 110 August SST forecasts were correlated (reference period: 1993 to 2009) to four observational references (ORs, legend). For each grid point, we display the OR that correlates best with most of the 110 forecasts. We show data only in grid points where at least one forecast achieves a significant correlation with one OR. The black box is the Niño3.4 region. Regions above 70°N or below 70°S are not considered, as these regions are usually ice-covered. The percentages below the color bar indicate the fraction of the oceans covered by each color. The corresponding figures for other months and alternative metrics can be found in the supplementary materials.

The differences in skill for different ORs, as depicted in Fig. 3, might be interpreted as an accidental result. Indeed, all forecasts are highly correlated to each other (fig. S2). If, by chance, the first forecast of the first model correlates better with the ESA-CCI OR, then it is likely that all subsequent forecasts would also tend to correlate better with that particular OR. We show, however, using bootstrapping and parametric tests, that a result as extreme as the one shown in Fig. 3 is very unlikely to have occurred by chance (*24*). In addition, the better performance of ESA-CCI in the Niño3.4 box is confirmed for another widely used metric of performance, the root mean squared error (RMSE) (*25*) (fig. S3). Finally, to push the analysis even further, this exercise is repeated on a different test case: summer Arctic sea ice extent prediction. In line with SST analyses, results show that sea ice concentration products relying on more recent algorithms tend to provide better estimates of model performance, or presented the other way around, that better observations are generally of higher skill when assessed against available forecasts (see supplementary materials for details).

The reasons why ESA-CCI outperforms other ORs in the Niño3.4 box are difficult to trace back formally. Indeed, the lack of information on SST uncertainty in all products, except ESA-CCI itself, precludes a clean comparison between the products' quality (*26*). It is therefore not possible to confirm whether the ESA-CCI error variance is lower than that of other products, as we would expect from lessons learned from the statistical toy model. However, we can gain further understanding of the role of observational quality if we extend the diagnostic of Fig. 3 to other regions of the world (Fig. 4). The map suggests that it is meaningless to designate any best OR in an absolute sense. Rather, it is clearly noticeable that where ORs are of a priori better quality, they score better: ORs that are partially or fully based on in situ data, shown in shades of blue, are more skillful in areas of intense shipping such as the North Atlantic (United States to Europe), the North Pacific (United States to Japan and China), or the Eastern Pacific (North America to South America).

The central idea behind our contribution relies on the symmetrical roles played by climate models and observational references as imperfect, but complementary, sources of information about reality. Symmetry enables one to revisit the concept of climate model evaluation and to expand it so as to simultaneously estimate the quality of observational references. A corollary is that ORs of higher quality will yield better skill scores to climate models, with systematic and non-negligible impacts on the estimated model performance. From the modeling side, considering multiple ORs in future model evaluation exercises such as CMIP6 (*27*, *28*) appears therefore as a priority. From the observational side, quantification of uncertainties will be vital to enable intercomparison of ORs and to account for observational uncertainty in model evaluation. In both cases, considering climate model evaluation as a bidirectional exercise (and not unidirectional as assumed by many up to now) is essential to remember that observations, no matter how good they appear, are also intrinsically uncertain.

REFERENCES AND NOTES

1. T. F. Stocker *et al.*, Eds, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, 2013).
2. World Meteorological Organization (WMO)'s Global Observing System for Climate (GCOS, www.wmo.int/pages/prog/gcos/) has defined a list of "essential climate variables" (ECVs) that are intended to best support the work of the United Nations Framework Convention on Climate Change (UNFCCC) and the Intergovernmental Panel on Climate Change (IPCC). Wind speed and direction, pressure, precipitation, and greenhouse gases, but also SST and sea ice, on which the present study is based, are primary ECVs.
3. G. Hegerl, *Environ. Res. Lett.* **10**, 071001 (2015).
4. M. Collins *et al.*, *Nat. Clim. Chang.* **2**, 403–409 (2012).
5. See, for instance, NASA's Earth Observing System (http://eospso.nasa.gov/), European Commission's Copernicus (www.copernicus.eu/), or China's Observation Satellite Programs (*29*).
6. There are protocols to assess the quality of ORs using independent in situ data; see, e.g., the GAIA-CLIM project www.gaia-clim.eu/. However, the data are by definition largely undersampled in space and time and cannot verify ORs exhaustively.
7. T. Reichler, J. Kim, *Bull. Am. Meteorol. Soc.* **89**, 303–311 (2008).
8. P. J. Gleckler, K. E. Taylor, C. Doutriaux, *J. Geophys. Res.* **113**, D06104 (2008).
9. A. P. Weigel, M. A. Liniger, C. Appenzeller, *Q. J. R. Meteorol. Soc.* **134**, 241–260 (2008).
10. B. Sanderson, R. Knutti, *Geophys. Res. Lett.* **39**, L1670 (2012).
11. S. Siegert *et al.*, *J. Clim.* **29**, 995–1012 (2016).
12. O. Bellprat, F. J. Doblas-Reyes, *Geophys. Res. Lett.* **43**, 2158–2164 (2016).
13. P. A. Janssen, S. Abdalla, H. Hersbach, J.-R. Bidlot, *J. Atmos. Ocean. Technol.* **24**, 1665–1677 (2007).
14. E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability* (Cambridge Univ. Press, 2002).
15. Linear correlation of variables X and Y is defined as $\rho = \frac{C(X,Y)}{\sqrt{V(X)\cdot V(Y)}}$, where C and V denote covariance and variance, respectively. Sample correlation, denoted r throughout the text, is calculated using the same formula but with the covariance and variance replaced by sample covariance and sample variance, respectively. For application in climate sciences, see, for example, (*30*).
16. A. Scaife *et al.*, *Geophys. Res. Lett.* **41**, 2514–2519 (2014).
17. H. Bellenger, E. Guilyardi, J. Leloup, M. Lengaigne, J. Vialard, *Clim. Dyn.* **42**, 1999–2018 (2014).
18. F. Massonnet *et al.*, *Cryosphere* **5**, 687–699 (2011).
19. The toy model introduced here has the advantage of generalizing other simple models used in the literature to mimic large multimodel ensembles. See (*11*), section 2b, and the supplementary materials for a thorough discussion about hypotheses and interpretation of this toy model.
20. T. Palmer, R. Hagedorn, Eds, *Predictability of Weather and Climate* (Cambridge Univ. Press, New York, 2014).
21. C. J. Merchant *et al.*, *Geosci. Data J.* **1**, 179–191 (2014).
22. In fact, the ESA-CCI product was released after the simulations were conducted. See (*21*) for a description of the ESA-CCI product.
23. The spatial availability of SST in situ data (and its limited sampling of data in the equatorial Pacific compared to other regions) can be explored in the following reference: C. Deser and National Center for Atmospheric Research Staff, Eds, "The Climate Data Guide: ICOADS Surface Marine Weather Observations" (https://climatedataguide.ucar.edu/climate-data/icoads-surface-marine-weather-observations).
24. We challenge the null hypothesis that correlations are identical for all observational products using bootstrapping and parametric tests. (i) Bootstrapping: Synthetic 1993–2009 time series for the 110 August forecasts and the four ORs are generated 100,000 times from the known sample covariance matrix of the data, which we modify so that for each forecast, the correlation with all four ORs is forced to be identical (see supplementary materials for a detailed explanation). By doing so, we generate artificial data with exactly the same statistical properties as those of the original sample, except that no OR is chosen to be a priori better than any other one. A situation as extreme as the one witnessed—i.e., that one OR chosen a priori ranks first for at least 93 forecasts out of 110 (Fig. 3)—occurs 1.9% of the time. That is, the null hypothesis that all observations correlate equally to the 110 forecasts is very unlikely to be true, even if the strong correlation among forecasts is accounted for. (ii) Parametric test: The test for an increase in correlation in the presence of nonindependent samples (*31*) is performed for the pair of ERSST4/ESA-CCI for the August forecasts. For 101 of the 110 forecasts (92%), the P value returned by the test is lower than 10% (84% have a P value lower than 5%), i.e., far from the expected uniform distribution that would prevail if there were no true difference in the correlations.
25. $RMSE(X,Y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2}$ where N is the number of available samples. Note that, like the correlation (*15*), the RMSE is a symmetric operator. All results obtained with the RMSE instead of the correlation are available in the supplementary materials.
26. Even if the grid-point information on daily SST uncertainty is available (as is the case for the ESA-CCI product), propagating the statistics to climate scales (regional and monthly means) is far from trivial, as the spatial and temporal decorrelation time scales of errors are not properly characterized.
27. World Climate Research Programme (WCRP)'s Coupled Models Intercomparison Project, phase 6: www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6.
28. V. Eyring *et al.*, *Geosci. Model Dev.* **9**, 1937–1958 (2016).
29. G. Xingfa, T. Xudong, *IEEE Trans. Geosci. Remote Sens.* **3**, 113–129 (2015).
30. K. E. Taylor, *J. Geophys. Res.* **106**, 7183–7192 (2001).
31. J. H. Steiger, J. M. Lind, "Statistically based tests for the number of common factors," paper presented at the annual meeting of the Psychometric Society, Iowa City, IA, 30 May 1980.
32. D. Wilks, *Statistical Methods in the Atmospheric Sciences* (Elsevier, Oxford, 2011), pp. 182–184.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the Belgian Fonds National de la Recherche Scientifique (FNRS), the Ministerio de Economía y Competitividad (MINECO) Juan de la Cierva program, the Spanish national project PICA-ICE (GA CGL2012-31987), the European Union Seventh Framework Programme FP7 projects SPECS (GA 308378) and EUCLEIA (GA 607085), and the European Space Agency Living Planet Fellowship Programme under the project VERITAS, as well as the Climate Model User Group (CMUG) of the European Space Agency's Climate Change Initiative. We thank the National Oceanic and Atmospheric Administration, NSF, NASA, and U.S. Department of Energy for providing access to the North American Multi-Model Ensemble (NMME) database, and we acknowledge the help of the National Centers for Environmental Prediction, International Research Institute, and National Center for Atmospheric Research personnel in creating, updating, and maintaining the NMME archive. Acknowledgment is made for the use of the European Centre for Medium-Range Weather Forecasts computing and archive facilities in this research. We gratefully acknowledge the computer resources, technical expertise, and assistance provided by the Red Española de Supercomputación (RES). All data used for the analyses conducted in this study are available at https://doi.org/10.1594/PANGAEA.864680. We acknowledge managers of the PANGAEA project for hosting this content.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/354/6311/452/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S5
Source Code
References (*33*–*42*)

20 April 2016; accepted 20 September 2016
Published online 6 October 2016
10.1126/science.aaf6369

Science
AAAS

Using climate models to estimate the quality of global observational data sets

François Massonnet, Omar Bellprat, Virginie Guemas and Francisco J. Doblas-Reyes (October 6, 2016)

Science **354** (6311), 452-455. [doi: 10.1126/science.aaf6369]
originally published online October 6, 2016

Editor's Summary

Models and data: A two-way street

Data are used to drive models of climate and other complex systems, but is the relationship between data and models a one-way process? Massonnet *et al*. used climate models to assess the quality of the observations that such models use. Starting with a simple model and progressing to more complex ones, the authors show that models are better when they are assessed against the most recent, most advanced, and most independent observational references. These findings should help to evaluate the quality of observational data sets and provide guidance for more objective data set selection.

Science, this issue p. 452

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.