nature climate change

SUPPLEMENTARY INFORMATION

DOI: 10.1038/NCLIMATE3041

Detecting failure of climate predictions

Michael C. Runge[1*], Julienne C. Stroeve[2], Andrew P. Barrett[2], Eve McDonald-Madden[3]

[1] USGS Patuxent Wildlife Research Center, Laurel, MD 20708 USA

[2] National Snow and Ice Data Center, University of Colorado, Boulder, CO 80309 USA

[3] School of Geography, Planning, and Environmental Management, University of Queensland, St. Lucia, QLD 4072 Australia

*Email mrunge@usgs.gov

Power analysis

The method proposed in this paper uses a fit statistic based on the empirical cumulative distribution to examine whether a model set bounds the truth. The power of the test to detect observed behavior that is outside the bounds of the model set is an important property for a decision-maker—such detection is the first step in the diagnostic and development process to improve the model set for future predictions. False detections, however, would trigger unnecessary effort to examine the existing model set. Thus, the decision maker will want to balance the Type I and Type II error rates in a manner fitting the particular decision context. For the particular type of test proposed (a sequential goodness-of-fit test to a best-weighted model with a moving window), the Type I and Type II error rates will be affected by the type of test, the length of the window, the degree of autocorrelation in the underlying process, the models in the proposed set, and the difference between the true process and the bounds of the model set. In

DOI: 10.1038/NCLIMATE3041

this Supplementary Information, the Type I and Type II error rates for the proposed method are explored.

Methods

Type I error rate. To examine the Type I error rate of the proposed method, 50-year time series data of annual observations were simulated from a generating model with a constant mean (54.5) and standard deviation (1.5), using a normal distribution. The Kolmogorov-Smirnov test statistic was calculated for the empirical distribution function for sequential 10-year windows, beginning at year 10 and incremented annually, compared to a proposed distribution that was the same as the generating model. For each of 10,000 replicates of this process, the maximum test statistic across the 41 sequential windows was calculated. The distribution of the maximum test statistic was used to calculate critical values that correct for the multiple comparisons across time.

Power analysis. To test the power of the proposed method to detect an underlying truth that is not bounded by the model set, three scenarios were simulated. The simulations all assumed annual observations over a 50 year period. The simulations were modeled loosely after the northern pintail case study presented in the main body of the paper.

Scenario 1. Two models, both normally distributed with constant mean and standard deviation, were compared: Model 1 predicted a constant mean of 53.5; Model 2 predicted a constant mean of 55.5; and both models predicted a standard deviation of 1.5. The generating model for the simulations was normally distributed with constant standard deviation of 1.5; the fixed mean for

the generating model varied between 52 and 57. For each set of values for the generating model, 100 replicates of the 50-year time series were simulated.

Scenario 2. Two models, both normally distributed with constant mean and standard deviation, were compared: Model 1 predicted a standard deviation of 1.1; Model 2 predicted a standard deviation of 1.9; and both models predicted a mean of 54.5. The generating model for the simulations was normally distributed with constant mean of 54.5; the fixed standard deviation for the generating model varied between 0.25 and 4.5. For each set of values for the generating model, 100 replicates of the 50-year time series were simulated.

Scenario 3. The model set consisted of the same two models as in Scenario 1. The generating model, however, had a time-dependent mean described by

$$\mu = 54.5 + \frac{2}{1+e^{-0.5(t-20)}} \qquad \text{(SI-1)}$$

which changes smoothly from 54.5 to 56.5 with the median change reached at $t = 20$ (Fig. SI-5A). The standard deviation of the generating model was 1.5. The time series was replicated 100 times.

Individual model fit. To assess the fit of each model to the data, a moving 10-yr window was used. Within the moving window, the observations were expressed as a normalized residual from the corresponding year-specific predicted distribution. An empirical cumulative distribution function was formed from the set of residuals within the window and compared

DOI: 10.1038/NCLIMATE3041

against the cumulative distribution function for a standard normal distribution to calculate the Kolmogorov-Smirnov statistic[1] or the Anderson-Darling statistic[2,3].

Weighted models. Weighted models were formed from the component models with linear weighting of the first two moments. The best-fit weighted model at each point in time was found by searching for the set of weights that minimized the goodness-of-fit statistic, using the fmincon function in MATLAB with the active-set optimization algorithm, which uses sequential quadratic programming[4].

Results

Type I error rate. The critical value (α=0.05) for a two-sided Kolmogorov-Smirnov test with a sample size n = 10 is 0.40925 (Table 1). Using 41 sequential moving 10-year windows of data for a 50-year time series, at least one of the 41 test statistics exceeds this critical value 54.1 percent of the time, rather than 5 percent of the time, because multiple comparisons are being made. (If the 41 windows were independent, at least one would exceed the nominal α = 0.05 critical value 87.8 percent of the time.) To fully control for these multiple comparisons, so that the Type I error rate of α = 0.05 applies family-wise to the series of 41 tests, the critical value should be 0.5466 (Table 1).

Table 1. Two-sided critical values for the Kolmogorov-Smirnov test with a sample size $n = 10$. For a test with a single 10-yr window, the exact critical values were found by solving the polynomial given by Miller[7]. For the test across 41 sequential 10-yr moving windows, the critical values were found by simulation.

	Critical values, two-sided				
α	**0.005**	**0.01**	**0.025**	**0.05**	**0.10**
Single 10-yr window	0.51872	0.48893	0.44562	0.40925	0.36866
Sequential 10-yr moving windows	0.6269	0.6075	0.5730	0.5466	0.5155

Power analysis. The proposed method readily detects true means outside of the bounds of the model set at some point in the sequence of 41 overlapping moving windows (Fig. SI-1). Using a nominal critical value, there is a fairly high false-positive rate of the true mean within, but near, the bounds of the model set (Fig. SI-1, black line). Using a critical value corrected for multiple comparisons (Table 1), the sensitivity of the test is reduced close to the model set boundary, but gains power the farther the true mean is from the bounds (Fig. SI-1, blue line). The Anderson-Darling statistic has a slightly lower false-positive rate at the boundary, and a slightly higher power outside the bounds of the model set, compared to the Kolmogorov-Smirnov statistic (Fig. SI-2).

DOI: 10.1038/NCLIMATE3041

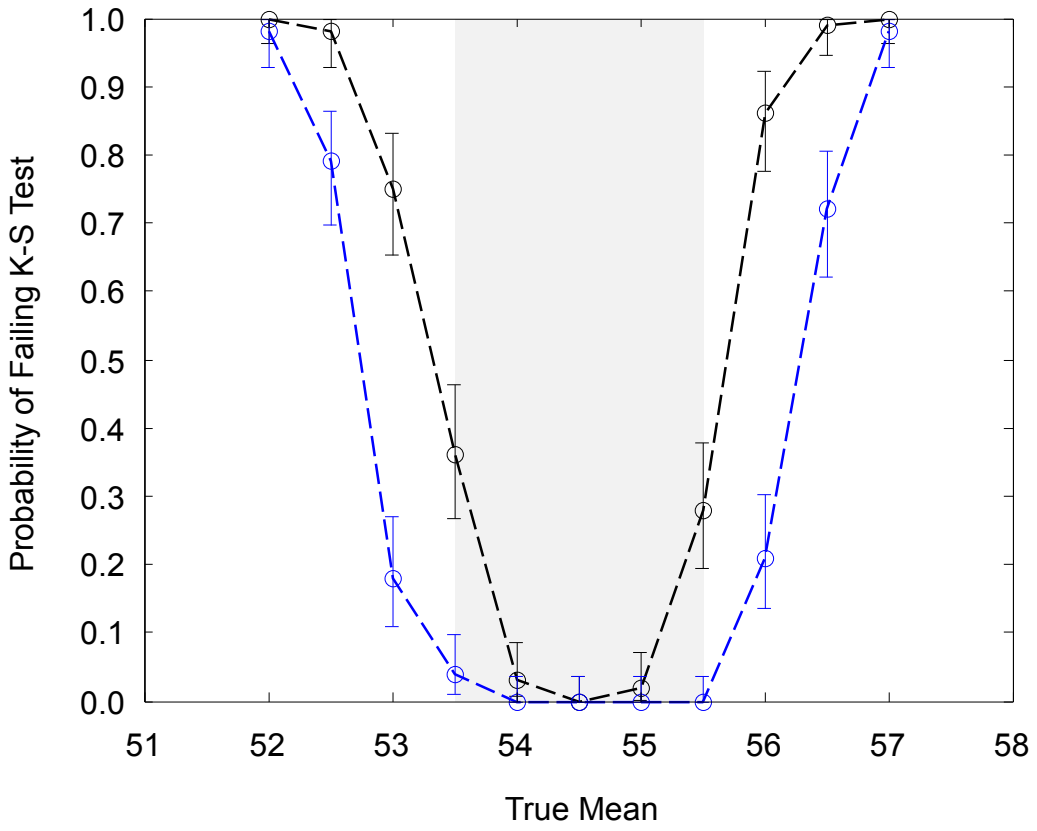

Figure SI-1. Power to detect a mean outside the bounds of the model set as a function of the true mean, with (blue) and without (black) correction for multiple comparisons. The data show the fraction of simulated replicates (n = 100) in which the Kolmogorov-Smirnov test of the best-weighted model exceeded an error rate ($\alpha = 0.05$) at some point during the 50 year simulation, with exact 95% binomial confidence intervals. The shaded region shows the bounds of the model set, which included two models ($\mu_1 = 53.5$, $\mu_2 = 55.5$; both models used $\sigma = 1.5$).

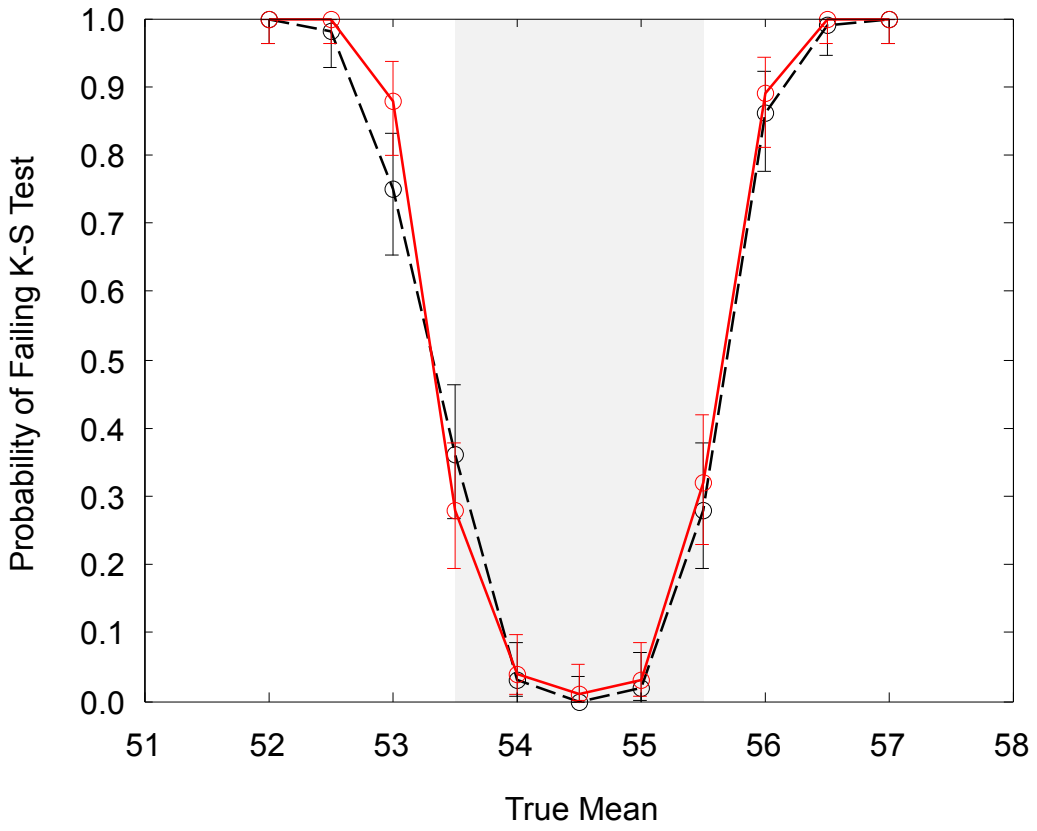

Figure SI-2. Power to detect a mean outside the bounds of the model set as a function of the true mean, comparing Kolmogorov-Smirnov (black) and Anderson-Darling (red) statistics. The data show the fraction of simulated replicates (n = 100) in which the goodness-of-fit test of the best-weighted model exceeded a nominal error rate (α = 0.05) at some point during the 50 year simulation, with exact 95% binomial confidence intervals. The shaded region shows the bounds of the model set, which included two models (μ_1 = 53.5, μ_2 = 55.5; both models used σ = 1.5).

The proposed test (using the Kolmogorov-Smirnov statistic) is less powerful at detecting a true standard deviation outside the bounds of the model set than it is at detecting a mean outside the bounds (Fig. SI-3). The nominal critical values produce a high false-positive rate, even for true standard deviations well within the bounds of the model set (Fig. SI-3, black line). When these critical values are corrected for multiple comparisons (Table 1), the Type I error rate is appropriate, but the power is greatly reduced, especially for true standard deviations that are

DOI: 10.1038/NCLIMATE3041

smaller than the values bounded by the model set (Fig. SI-3, blue line). The Anderson-Darling statistic is better than the Kolmogorov-Smirnov statistic for detecting departures in the standard deviation from the bounds of the model set, as it is both more specific and more sensitive (Fig. SI-4).

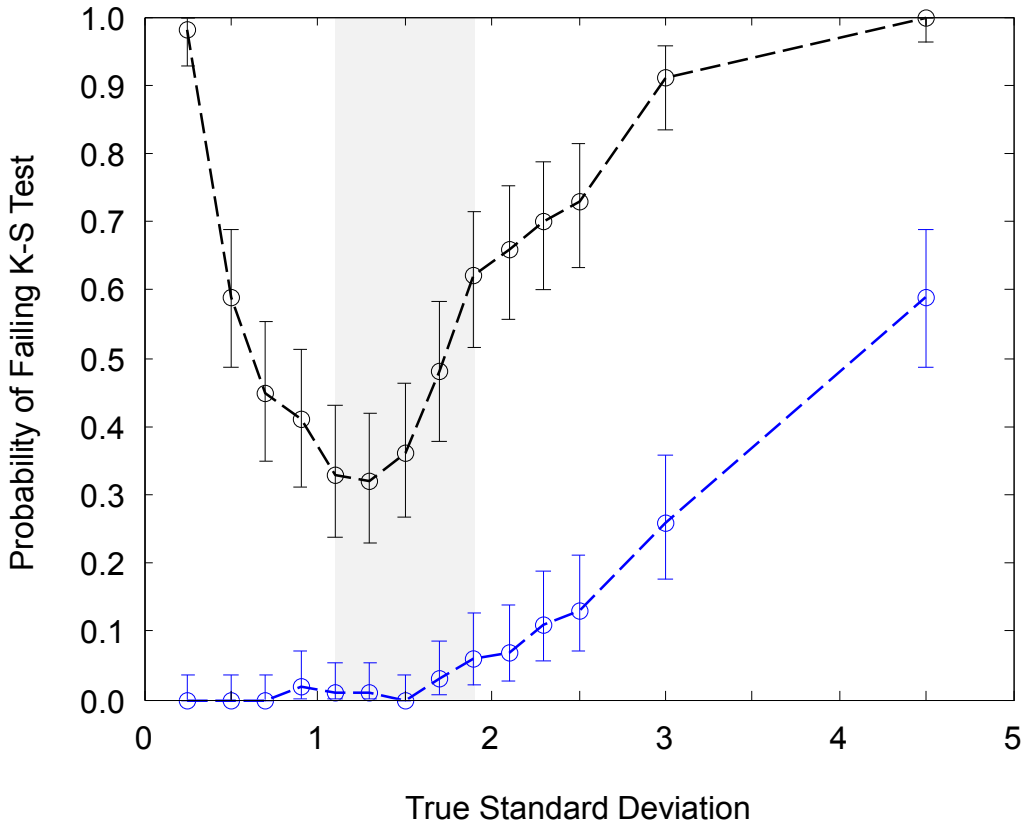

Figure SI-3. Power to detect a standard deviation outside the bounds of the model set as a function of the true standard deviation, with (blue) and without (black) correction for multiple comparisons. The data show the fraction of simulated replicates (n = 100) in which the Kolmogorov-Smirnov test of the best-weighted model exceeded an error rate ($\alpha = 0.05$) at some point during the 50 year simulation, with exact 95% binomial confidence intervals. The shaded region shows the bounds of the model set, which included two models ($\sigma_1 = 1.1$, $\sigma_2 = 1.9$; both models used $\mu = 54.5$).

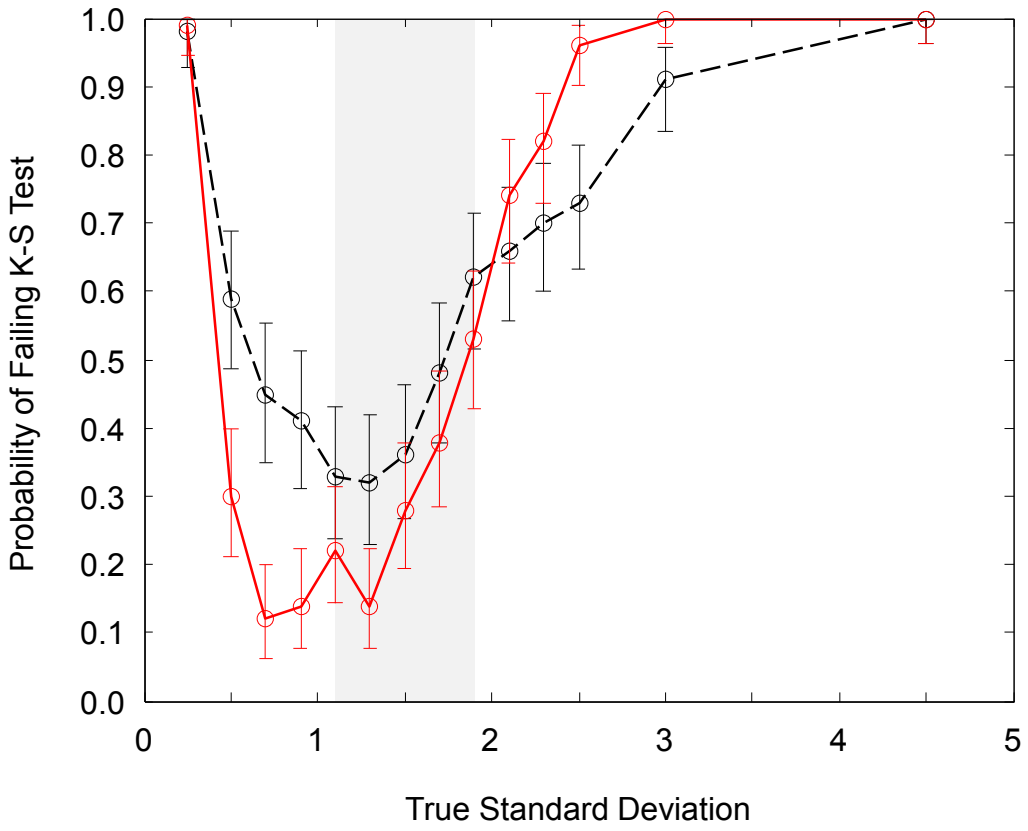

Figure SI-4. Power to detect a standard deviation outside the bounds of the model set as a function of the true standard deviation, comparing Kolmogorov-Smirnov (black) and Anderson-Darling (red) statistics. The data show the fraction of simulated replicates (n = 100) in which the goodness-of-fit test of the best-weighted model exceeded a nominal error rate ($\alpha = 0.05$) at some point during the 50 year simulation, with exact 95% binomial confidence intervals. The shaded region shows the bounds of the model set, which included two models ($\sigma_1 = 1.1$, $\sigma_2 = 1.9$; both models used $\mu = 54.5$).

In a simulated time series with a changing mean, the average weight on the two models remains around 0.5 when the true mean lies between them (Fig. SI-5B). In the first 10 years after the true mean leaves the bounds of the model set (years 20-30 in the simulation), the weight on model 1 (Fig. SI-5A, red dashed line) drops to nearly 0 (Fig. SI-5B). At the same time, the Kolmogorov-Smirnov test (using a nominal critical value) begins to detect failure of the model set (Fig. SI-5C). By year 30 (10 years after the true mean has moved outside of the bounds of the model set),

DOI: 10.1038/NCLIMATE3041

the proposed method detects a departure 34% of the time; by year 40, it detects a departure 84% of the time; by year 50, it detects a departure 93% of the time.

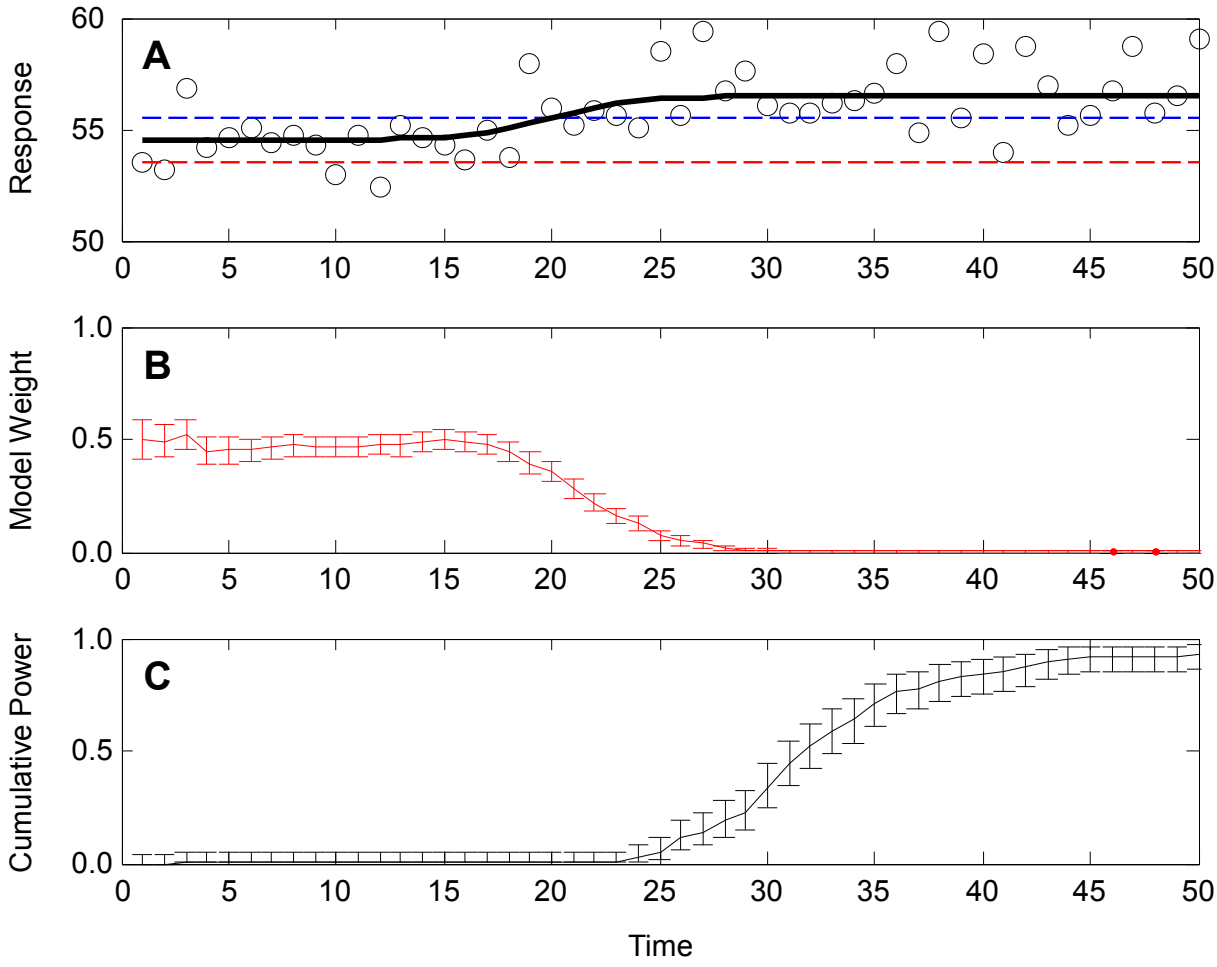

Figure SI-5. Power to detect a change in mean as a function of time. (A) Simulated observed data (single replicate). The true time-dependent mean is shown with the black line. The model set included two hypotheses ($\mu_1 = 53.5$ [red dashed line], $\mu_2 = 55.5$ [blue dashed line]; both models used $\sigma = 1.5$). (B) Mean weights (with 95% confidence intervals, $n = 100$ simulated replicates) on Model 1 (red) that provide the best fit to a 10-year moving window of observations, as measured by the Kolmogorov-Smirnov statistic. (C) Cumulative probability of concluding that the model set does not bound the truth, as determined by a Kolmogorov-Smirnov test with nominal error rate $\alpha = 0.05$.

Discussion

The method proposed in this paper does have the power to detect departures of the mean and standard deviation outside the bounds of the model set (Figs. SI-1 and SI-3); this power is increased with the use of the Anderson-Darling statistic compared to the Kolmogorov-Smirnov statistic (Figs. SI-2 and SI-4), as others have shown[5]. The nominal critical values produce a high false-positive rate because of multi-comparisons; the Type I error rate can be corrected, but it comes at the expense of the power of the test (Figs. SI-1 and SI-3).

Another subtle issue regarding calculation of the critical values is whether any of the parameters of the forecast have been estimated from the data used to test goodness of fit[6]. For example, in the pintail example in the main body of the paper, observations in years 1981-1985 are used to estimate the mean of Model 2. Thus, the EDF tests for the windows that include any of the years 1981-1985 should technically account for the parameter estimation in the determination of the critical values. For the demonstration purposes of this paper, this consideration is not of paramount importance and we have not undertaken the appropriate correction, but there may be settings where this effect needs to be accounted for.

In the situation where the underlying true dynamics are changing over time, it is worth noting that the weights on the models begin changing before the test starts detecting departures from the model set (Fig. SI-5). Thus, the weights are an indicator of system change, and might also be a useful early indicator of the failure of the model set.

DOI: 10.1038/NCLIMATE3041

In a decision-making context, the key question is how to balance the Type I and Type II error rates: the decision maker will want to detect departures from the model set as early as possible, but will not want to invest unnecessarily in the expense of model diagnosis and development. One approach would be to carefully weigh the costs of Type I and Type II errors and set an appropriate critical value (taking account of multiple comparisons). Another approach is to use a nominal critical value that produces a high false-positive rate, but use any test failure simply as an opportunity to inspect the time series and model performance, not necessarily as a commitment to a full effort in model revision. The details of how to set and manage the error rates are value judgments that depend on the specifics of the decision-making context.

References

1 **Massey, F. J., Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association* 46, 68-78, doi:10.2307/2280095 (1951).**

2 **Anderson, T. W. & Darling, D. A. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* 23, 193-212 (1952).**

3 **Anderson, T. W. & Darling, D. A. A test of goodness of fit. *Journal of the American Statistical Association* 49, 765-769 (1954).**

4 **Fletcher, R. *Practical Methods of Optimization*. (John Wiley and Sons, 1987).**

5 **Razali, N. M. & Wah, Y. B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics* 2, 21-33 (2011).**

6 **Babu, G. J. & Rao, C. R. Goodness-of-fit tests when parameters are estimated. *Sankhyā: The Indian Journal of Statistics* 66, 63-74 (2004).**

7 **Miller, L. H. Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association* 51, 111-121 (1956).**