# Beyond the Means: Validating Climate Models with Higher Order Statistics

David W. Pierce

*Climate Research Division, Scripps Institution of Oceanography*

*La Jolla, California*

*dpierce@ucsd.edu*

*Version #2: 20 January 2004*

*Submitted to Computing in Science and Engineering*

***Keywords: Earth and atmospheric sciences; model validation and analysis; mathematics and statistics***

# ABSTRACT

Large-scale climate models are validated by comparing the model's mean and variability to observations. New applications are placing more demands on such models, which can be addressed by examining the models' distributions of daily quantities such as temperature and precipitation.

What determines the climate where you live? Why does it vary, so that some years have unusually cold winters, or particularly hot summers? What will next winter be like? What will the climate be in coming years, and is it affected by emissions of gasses such as $CO_2$?

These are just a few of the questions that are examined with coupled ocean-atmosphere general circulation models (O-A GCMs). Such models are complicated, incorporating the equations of motion for air and water masses, the properties of sea ice, parameterizations for cloud processes, schemes for river flow, and the effects of soil moisture and ground cover. The projections given by such models might influence decisions ranging from whether someone's aging roof should be repaired before the coming winter to what future technologies the automobile industry should pursue. How are such models validated, so that we understand what confidence should be placed in their predictions?

This is typically done by comparing the model's behavior to that of the real world. The assumption (usually implicit) is that if the two behave similarly over some validation time period, then the model is likely to give skillful predictions of the future. This assumption is examined below, but the main focus here is how to compare models to the real world. The enormous range of space and time scales over which the atmosphere and oceans can vary makes this a surprisingly difficult question. Simply describing all this variability is a complicated task, much less comparing it to the limited set of observations that exist.

## The meaning of means

Model validations generally begin by comparing the model's monthly or seasonal mean fields to the observations. This kind of validation goes back to the earliest days of numerical ocean and atmosphere modeling; for example, it is used to good effect in the large body of work by S. Manabe[1] and the innovative ocean modeling of K. Bryan[2]. In practice, there is no objective standard for "how close" a model's mean monthly fields should be to observations before the model is used. Often modelers find that altering a parameterization (of cloud processes, for example) in order to improve agreement between the model and observations at one location, or at a particular time of the year, will decrease the agreement elsewhere or at other times. Different models have different strengths and weaknesses, so there is usually not one obviously "best" model to be used for all tasks.

A few techniques are used to address the problem of optimizing the model performance when changes can have both good and bad effects. First, climate modelers from many institutions came together to support the Atmospheric Model Intercomparison Project[3] (AMIP), which systematically compared atmospheric model fields from a large number of different models. The ability to see how different model configurations affected the simulations (and the incentives provided by having a well-performing model in an open competition!) led to improvements in many models, advancing the whole field. Second, the results from different models can be averaged together into a "multi-model ensemble." The idea is that any one model's bad aspects will be minimized by incorporating data from other models that do a better job in that area. There is some evidence that this can improve the quality of forecasts[4].

Much of the mean state of our climate is a determined by a strong radiative balance. The sun shines on the earth (incoming shortwave radiation), the warm atmosphere emits longwave radiation both downwards towards the surface and out into space, and the surface's temperature increases until it re-emits as much heat as it receives. It is relatively small departures from this mean balance that lead to the interesting climate variability we are generally concerned with.

## Climate variability

Behavior that depends on small perturbations to large, competing effects can be difficult to capture in a model, but are often the behaviors of interest in the climate system. For example, Fig. 1 shows an O-A GCM's estimated evolution of the surface heat flux components driven by increases in atmospheric $CO_2$ over the past 120 years. The top panel shows the actual values (globally and annually averaged), while the bottom panel shows the departure from conditions averaged between 1880 and 1930. The terms are on the order of 300 $W/m^2$. The *changes* in the upward and downward longwave radiation terms are on the order of 5 $W/m^2$, and these nearly compensate to give a *net* change in surface heat flux of only 1 $W/m^2$.

One way to probe how well a model captures small departures from the large mean state is to compare the model's variability to observations. Since natural variability in the climate is affected by how competing effects conspire to produce the mean, comparing a model's variability to that observed is a more sensitive test of model quality than just comparing the means.

For example, Fig. 2 shows a simple model of the climate whose surface temperature is determined by the balance between incoming radiation (sunlight plus thermal radiation from the atmosphere) and outgoing thermal radiation from the surface. Clouds have competing effects on the climate that are not completely understood, and so are parameterized in models. The blue lines in Fig. 2 show two hypothetical models with different cloud parameterizations. In model M-LW, the downwards thermal radiation clouds emit dominates as surface temperatures increase; in model M-R, the way clouds reflect sunlight dominates[5]. Both models have the correct surface temperature, $T_s$ (the intersection of the curves in the left panel of Fig. 2). Then imagine dust from a volcanic eruption decreases incoming sunlight. As shown in the right panel of Fig 2, this perturbation leads to different new surface temperatures for the two models, despite the fact that they started with the same mean and experienced the same perturbation.

It is generally assumed that a model with an accurate mean climate *and* realistic climate variability will make better predictions, but this presupposes that future perturbations will be the same kind as exist in natural variability. For example, in Fig. 2 again, imagine that with better understanding models M-LW and M-R are improved until they converge to the same line. With the same mean and natural variability, they would be considered equally validated by this method. If a new pollutant with effects unlike that involved in natural variability is added to the atmosphere, it is entirely possible that the two models will still give different predictions. ($CO_2$ exists and fluctuates naturally in the atmosphere, and so probably does not qualify as a "new" pollutant whose effects cannot be validated this way, but some human-generated aerosols from industry and burning might[6]).

The climate modeling community recognized the importance and difficulty of the validation question in the 1980s, as the field moved from more theoretical and abstract representations of the Earth to more detailed ones and applied questions. One result of this was the establishment, in 1989, of the Program for Climate Model Diagnosis and Intercomparison (PCMDI; http://www-pcmdi.llnl.gov), which brings analysis tools together with observed and model data. An interesting example of the analysis techniques PCMDI has produced is the Taylor diagram[7], which shows multiple aspects of a model/observation comparison simultaneously (http://www-pcmdi.llnl.gov/pcmdi/pubs/pdf/55.pdf).

A new emphasis on more subtle aspects of climate variability has been driven by climate change research. This seeks to identify how the climate may be altered due to human activity, such as the release of $CO_2$ into the atmosphere from smokestacks and tailpipes. One interesting finding is that some of the observed increase in global temperatures in recent decades is due more to warmer daily minimum temperatures than daily maximum temperatures[8]. In other words, the daily range of temperatures is decreasing and the distribution of hourly temperatures is changing. Another interesting result is that changes in the frequency of extreme climate events, such as "hot" days or "heavy" rainfall, may be one of the most important outcomes of this process[9].

Both these results suggest that a finer examination of the actual daily *distributions* of climate variables, how well they are simulated and how they might change in the future, is the next logical step in the progression of O-A GCM validation. This is made possible by the increasing quality of models, which are doing better than ever in simulating climate variable means and variance. There are also other important applications besides climate change that are interested in distributions. For example, people tend to turn on their air conditioners when the temperature gets warm enough, so the distribution of summertime daily maximum temperatures is of interest to energy producers, while daily precipitation is relevant to flooding. How much confidence can we place in the distributions such models evolve? Are they similar to what is observed?

## Climate variable distributions

As you might expect, models (and nature) show a wide variety of distributions for interesting climate variables, depending on the variable, the location, and the time of year. For example, Fig. 3 shows histograms of an O-A GCM's daily winter precipitation and average temperature at Ontario, Canada, and cloud cover in the southwestern U.S. The precipitation histogram is highly skewed, while temperatures follow a more Gaussian distribution (but see below). Cloud cover in the model tends to be almost binary, with many clear days, some completely cloudy days, and a small, almost constant occurrence of days between these extremes.

Clearly, describing *all* these fields simply in terms of a mean and variance can be misleading. For example, the mean cloud cover value of 0.42 is of little worth in describing a situation where days are typically either completely clear or completely cloudy, and it is quite unusual to have a day with fractional cloudiness between 0.4 and 0.5.

One approach to this problem of compactly, but more accurately, describing climate variability is to choose appropriate theoretical distributions for each variable, then present the parameters for the best-fit curve[10]. For example, temperature could be modeled by a Gaussian distribution, and the two controlling parameters (mean and standard deviation) reported. Precipitation could be modeled by a gamma distribution, and the shape and scale parameters retained, while cloudiness could be modeled by a two-parameter beta distribution. By way of illustration, the best-fit curves for the appropriate distribution are show in in Fig. 3 for each variable.

Athough this approach is not currently used in O-A GCMS, it is perfectly practical, and computationally efficient in the following sense. Many distributions useful for climate variables, including the Gaussian, gamma, two-parameter beta, and lognormal can be fit to a set of data using only two accumulations (partial sums) based on the data. For example, the Gaussian and beta distributions can be fit given $\Sigma x_t$ and $\Sigma x_t^2$ (where $x_t$ is the data at time t); the gamma can be fit given $\Sigma x_t$ and $\Sigma \ln(x_t)$; and the lognormal can be fit given $\Sigma \ln(x_t)$ and $\Sigma \ln(x_t)^2$. Typically, climate models are run saving averaged *monthly* output, for which purpose they use an accumulation $\Sigma x_t$. With only twice as much storage for each variable of interest, the proper accumulations can be retained to compute, during runtime, the best-fit distributions for *daily* climate variables. Otherwise, to calculate the best-fit daily distributions directly would require saving the daily data (~30 times as much output as saving monthly data) and post-processing, which in practice becomes a data volume and rate that is difficult to deal with.

## Distributions of temperature and precipitation in the contiguous U.S.

So how well does a modern O-A GCM actually do in capturing the distribution of daily temperatures and precipitation in the U.S.? For that matter, what do the observed distributions look like – are they well fit by the theoretical distributions suggested above?

These questions can be answered using daily observations of temperature and precipitation over the contiguous U.S.[11] Let’s step back a moment and see whether the Gaussian distribution suggested above is actually appropriate for temperature (temperature is examined here because there is a fair body of literature already comparing precipitation to gamma distributions, but the distribution of daily temperature has received little attention despite its importance). One way of doing this is to make use of the skew (*s*) and kurtosis (*k*) of the sample data, defined as:

$$s = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3$$

$$k = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4$$

where $x_i$ is the time series of data, *n* is the number of observations, $\bar{x}$ is the sample data's mean, and σ is the standard deviation. Such higher order statistics should be used cautiously, since they can be hard to estimate accurately from small samples of data. Here, we'll use 10 years of daily data from a particular 3-month season, which yields ~300 independent data points (taking account of the serial autocorrelation time of a few days).

The values of *s* and *k* are unbounded, so are not convenient to plot. This can be surmounted by mapping *s* and *k* to the unit plane using a so-called $\Theta_1$-$\Theta_2$ representation[12]. This representation discards the sign of the skew parameter, however, which is interesting to retain – are daily temperatures positively skewed or negatively skewed? – so instead we will use the following transformation:

$$\Theta_s = \frac{s}{1+|s|}$$

$$\Theta_k = \frac{1}{k}$$

With this transformation the theoretically expected values for many common distributions fall in the region shown in Fig. 4, and can take the form of points, lines, or regions, depending on how flexible the distribution is. For example, the points associated with the Gaussian, uniform, and exponential distributions are shown, along with the line of the gamma distribution and the region of the beta distribution.

Overplotted in the left panel of Fig. 4 are observed values of $\Theta_s$ and $\Theta_k$ from weather stations in the contiguous U.S., for the average daily wintertime (December, January, and March) temperature *anomalies* during the decade of the 1990s. (Anomalies are departures from mean conditions for that time of year, and are used to remove the strong effect of the annual cycle.) The values tend to have consistent negative skew, unlike a true Gaussian distribution. Of course, there will be sampling fluctuations in the estimates due to the finite size of the sample used; the heavy dashed line shows the region expected to enclose 95% of the data points. Many more data points fall outside this region than should, were the underlying distribution actually a Gaussian. Both these suggest that U.S. daily average temperature anomalies, in winter, tend to have a non-Gaussian distribution, with a non-negligible negative skew.

How well does a modern O-A GCM capture this behavior? The right panel of Fig. 4 shows values from the Parallel Coupled Model[13], a state-of-the-art climate model run with an atmospheric resolution of approximately 2.8° in latitude/longitude and 18 vertical levels, an ocean model with about 2/3° horizontal resolution and 32 vertical levels, a sea ice model that includes dynamics and thermodynamics, and a simple land cover scheme that includes vegetation, soil moisture, and runoff. Note that there are less points for the model than for observations because the model's resolution provides less measurements over the U.S. than there are observation stations. The model tends to generate average daily temperatures that have a wider range of skew than observed, as well as having a larger $\Theta_k$ (i.e., smaller kurtosis, or flatter distribution).

What about other seasons? Fig. 5 shows similar plots for data during northern hemisphere summer (June-July-August). The observations show many more stations with positive skew than are seen in winter, although there is still a tail of stations with negative skew. The model results for summer are little changed from the winter results, suggesting that the model does not capture the change between summer and winter processes as well as it might.

It is also interesting to look at how the characteristics of the distributions vary with location. Fig. 6 shows local $\Theta_s$ for average daily temperature anomalies in winter, for both the observations and the O-A

GCM. The model gets some of the large scale features correct without capturing the details; for example, the most negatively skewed values tend to occur in the mountainous regions of the Northwest as well as the southern part of Florida. The model reproduces this but also extends the tongue of negative skew values too far into the central plains states. The regions of most positive skew are found in mid to southern California and a region centered on Louisiana; the model has a local maximum of skew in Southern California, but in the Gulf states the a maximum is displaced noticeably to the west, falling in Texas.

We saw by comparing Fig. 4 to Fig. 5 that the skew for temperature anomalies was rather different between the summer and winter in the observations, but less so in the model. Figure 7 shows the geographical distributions for summer conditions. As anticipated, large differences between summer and winter conditions can be seen in the observations (compare to Fig. 6); the upper Midwest and Pacific Northwest coastal regions change to positively skewed values in the summer, while the Gulf coast region around Louisiana changes to negatively skewed conditions. The summer-winter difference in the O-A GCM and correspondence with observations are more modest, indicating that processes determining daily average temperature are better captured in winter than in summer. For example, the model misses the fringe of positively skewed stations along the west coast in the summer.

## The Larger View

It is interesting to think about some of the larger issues this approach touches on. For example, what have we discovered about the quality of the model? How applicable are these techniques to other problems, or other modeling approaches of the same problem? And can we say how much model validation is "enough"?

A useful way of approaching these issues is to realize that O-A GCMs are *tools* intended to address particular *questions*. The judge of quality, then, lies in how well the model simulates aspects of the system relevant to the question involved. For example, we saw in Fig. 7 that the model misses the fringe of positively skewed daily temperature distributions along the west coast. Histograms in this region show that the model eliminates the tail of very hot days from the observed distribution. Summer electricity use in that region is largely driven by those hot days, because of air conditioning loads. So, this particular model would not be a good tool to examine summer electricity loads on the U.S. west coast. On the other hand, the shape of the summer temperature distributions in the New England states is close to that observed; looking at electricity use in that region with this model would be fully justified. Other models might well have different locations where they perform well or poorly.

The key idea is that the answer to a question or application depends on particular variables and aspects of those variables (mean, variance, skew, etc.). A model's simulation of those can be rigorously characterized, using techniques such as described here. The experimenter then has to evaluate whether the simulation is realistic enough to help answer the question at hand. If so, then the validation may be thought of as "sufficient," for that model and that question.

This is a very broad, top-level view of a validation process, and one that is applicable to a wide variety of fields. It has particular emphasis in climate research is because there is only one system (the Earth), and controlled experiments on the Earth's climate are not possible. As a result, simulations are the only viable means of experiment, and it must be assumed that a simulation that accurately reproduces observed means, variances, skews, and other statistical measures does so because it is realistically simulating the actual causal physics involved.

We've been talking here about O-A GCMs, which are dynamical models that integrate the equations of motion forward in time, typically on a grid discretized in space or in frequency (for the spectral approaches). There is another whole class of climate models –statistical models -- that start with the basic idea that the simulated statistics *must* be correct in order to have a useful climate model. (By contrast, O-A GCMs start with the basic idea that the simulated physics must be correct for a useful model, then work forward to the statistics). A statistical model might, for example, use multiple regression to predict temperatures at an uninstrumented location based on temperatures at nearby weather stations; or it might use monte-carlo techniques to generate “synthetic weather” whose statistical characteristics mimic those observed. There are, of course, much more complicated and interesting types of statistical models[10].

While statistical models are extremely useful, the current fashion seems to be to use O-A GCMs for many of these purposes instead. There are several reasons for this. One is that it is trivial (although possibly time-consuming) to start with the dynamics encapsulated in an O-A GCM and empirically derive its statistics, merely by running the model for a long enough time. By contrast, it is far more difficult to start with the statistics of an observed variable and derive the form of the underlying dynamics. Another is that O-A GCMs naturally simulate the wave-like motions in the atmosphere that cause weather at locations separated in space or time to be linked. These linkages have to be built into statistical models explicitly, which means they must be known beforehand. This can be difficult in data-poor regions, when a large network of stations is considered, are when the remote linkages are not well understood.

## Conclusions

The preceding is meant to give a flavor for the kinds of analyses that can be done, and information obtained, by examining the *distributions* of model variables and comparing them to observations. There is much more than can be done with this technique; for example, by examining daily high or low temperatures (rather than just the average examined here), looking at other climate variables, or examining the kurtosis parameter rather than the skew. Distributions can be compared between El Nino years and La Nina years to see how global-scale climate fluctuations affect local weather. In addition to showing how the model simulates the system, these can reveal interesting aspects of the real world. For example, the skew of daily maximum temperatures differs from that of daily average temperature over the Northeast U.S., while the skew of daily minimum temperatures differs over the west. Ultimately, these kinds of comparisons can uncover reasons for why climate behaves as it does and how well models simulate that behavior.

This kind of validation of a complicated model is useful for situations where there is only one example (the Earth) and no possibility of controlled experiments. Moving from comparing means, to comparing variability, to (as proposed here) comparing the distributions of variables is motivated by a range of applications including climate change, flood control, and energy demand forecasting, and is made possible by the ever-increasing realism of large scale O-A GCMs.

## References

1. S. Manabe, J. Smagorinsky, and R. F. Strikler, "Simulated Climatology of a General Circulation Model with a Hydrologic Cycle," *MWR*, 769 (1965).

2. K. Bryan and L. J. Lewis, "Water Mass Model of the World Ocean," *JGR Oceans and Atmospheres,* **84**, 2503 (1979).

3. W. L. Gates, J. S. Boyle, C. Covey et al., "An Overview of the Results of the Atmospheric Model Intercomparison Project (Amip I)," *B Am Meteorol Soc,* **80** (1), 29 (1999).

4. C. Ziehmann, "Comparison of a Single-Model Eps with a Multi-Model Ensemble Consisting of a Few Operational Models," *Tellus A,* **52** (3), 280 (2000).

5. V. Ramanathan, R. D. Cess, E. F. Harrison et al., "Cloud-Radiative Forcing and Climate - Results from the Earth Radiation Budget Experiment," *Science,* **243** (4887), 57 (1989).

6. V. Ramanathan, P. J. Crutzen, J. T. Kiehl et al., "Atmosphere - Aerosols, Climate, and the Hydrological Cycle," *Science,* **294** (5549), 2119 (2001).

7. K. E. Taylor, "Summarizing Multiple Aspects of Model Performance in a Single Diagram," *JGR,* **106**, 7183 (2000).

8. D. R. Easterling, B. Horton, P. D. Jones et al., "Maximum and Minimum Temperature Trends for the Globe," *Science,* **277** (5324), 364 (1997).

9. G. A. Meehl, F. Zwiers, J. Evans et al., "Trends in Extreme Weather and Climate Events: Issues Related to Modeling Extremes in Projections of Future Climate Change," *B Am Meteorol Soc,* **81** (3), 427 (2000).

10. Daniel S. Wilks, *Statistical Methods in the Atmospheric Sciences : An Introduction*. (Academic Press, San Diego, 1995).

11. J. K. Eischeid, P. A. Pasteris, H. F. Diaz et al., "Creating a Serially Complete, National Daily Time Series of Temperature and Precipitation for the Western United States," *J Appl Meteorol,* **39** (9), 1580 (2000).

12. S. M. Abourizk and D. W. Halpin, "Statistical Properties of Construction Duration Data," *J Constr Eng M Asce,* **118** (3), 525 (1992).

13. W. M. Washington, J. W. Weatherly, G. A. Meehl et al., "Parallel Climate Model (Pcm) Control and Transient Simulations," *Clim Dynam,* **16** (10-11), 755 (2000).

## Figure Captions

1. Upper: components of the surface heat balance for an O-A GCM forced with historical $CO_2$ emissions over the period 1870-1999. LW-DOWN is downwards longwave (thermal) radiation from the atmosphere, "Short" is incoming shortwave (solar) radiation, LW-UP is upwards longwave from the surface, and NET is the sum of the other components. Lower: difference from 1880-1930 average.

2. Left: simple conceptual climate model, showing how the surface temperature is determined by the balance between the incoming and outgoing radiation. Two models with different cloud parameterizations are shown (M-R and M-LW). Right: after a uniform decrease to the incoming solar radiation.

3. Histograms of daily precipitation and temperature at Ontario, Canada, and daily cloud cover in the southwest U.S., from an O-A GCM.

4. A region of the $\Theta_s$-$\Theta_k$ plane, illustrating (transformed—see the text) skew and kurtosis parameters estimated from average daily temperature anomalies in winter (Dec-Jan-Feb) over the contiguous U.S. Left panel: black points show observed station data, 1990-1999. Right panel: black points show data from an O-A GCM. The theoretically expected values for Gaussian (red), exponential (blue), uniform (green), gamma (purple), and beta distributions (orange) are also shown.

5. As in Figure 4, but for summer (Jun-Jul-Aug).

6. Geographical distribution of transformed skew parameters for daily average temperature anomalies in winter (Dec-Jan-Feb). Top: observed station data, 1990-1999. Bottom: for an O-A GCM.

7. As in Figure 6, but for summer (Jun-Jul-Aug).

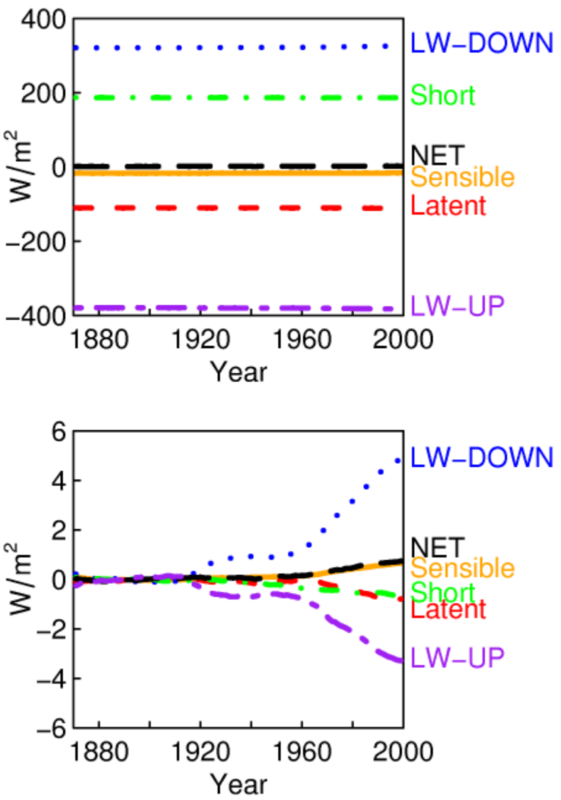

Figure 1

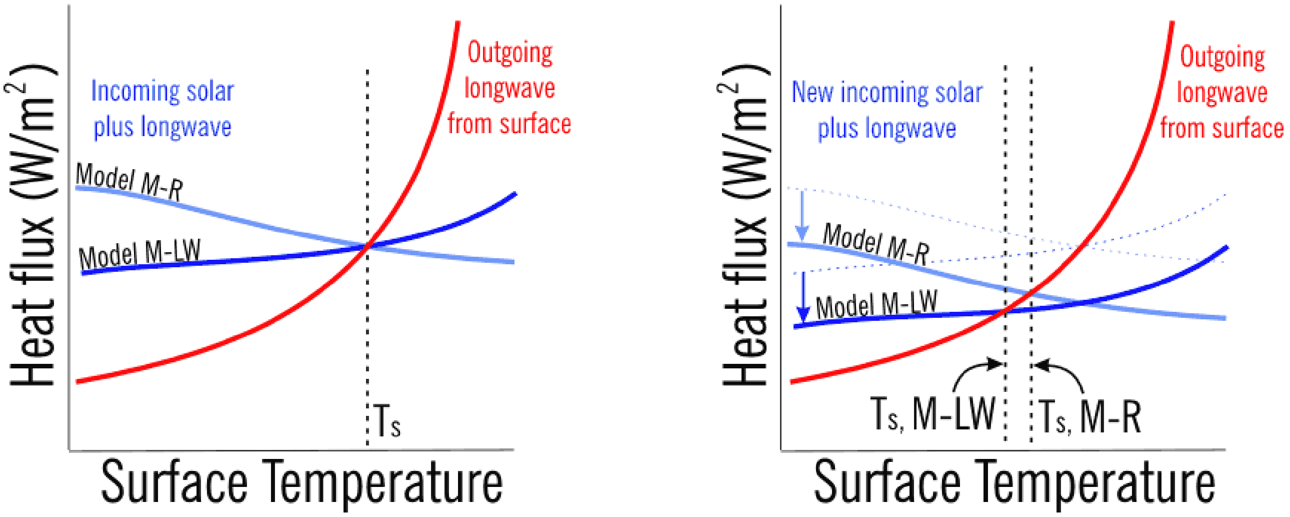

**Figure 2**

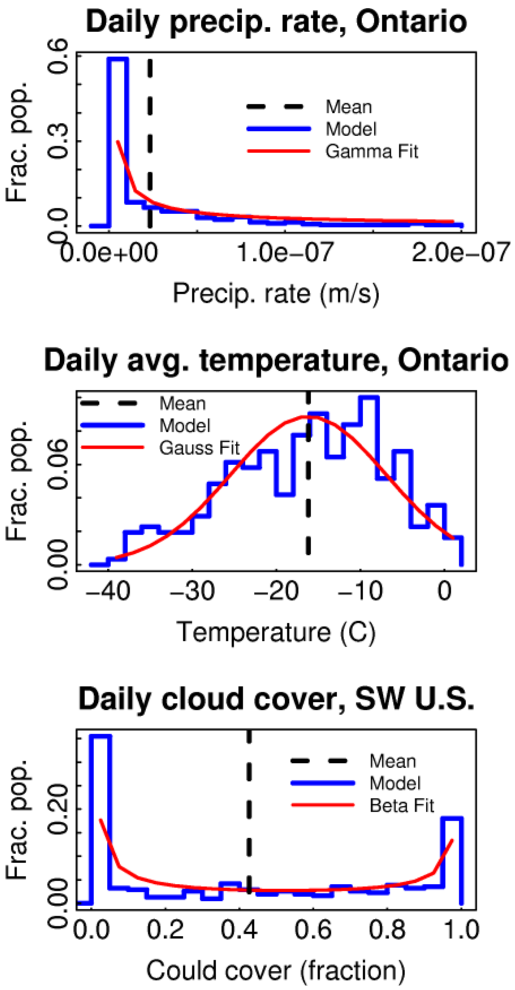
Daily precip. rate, Ontario
Frac. pop.
0.0
0.3
0.6
Mean
Model
Gamma Fit
0.0e+00
1.0e−07
2.0e−07
Precip. rate (m/s)
Daily avg. temperature, Ontario
Frac. pop.
0.00
0.06
Mean
Model
Gauss Fit
−40
−30
−20
−10
0
Temperature (C)
Daily cloud cover, SW U.S.
Frac. pop.
0.00
0.20
Mean
Model
Beta Fit
0.0
0.2
0.4
0.6
0.8
1.0
Could cover (fraction)

**Figure 3**

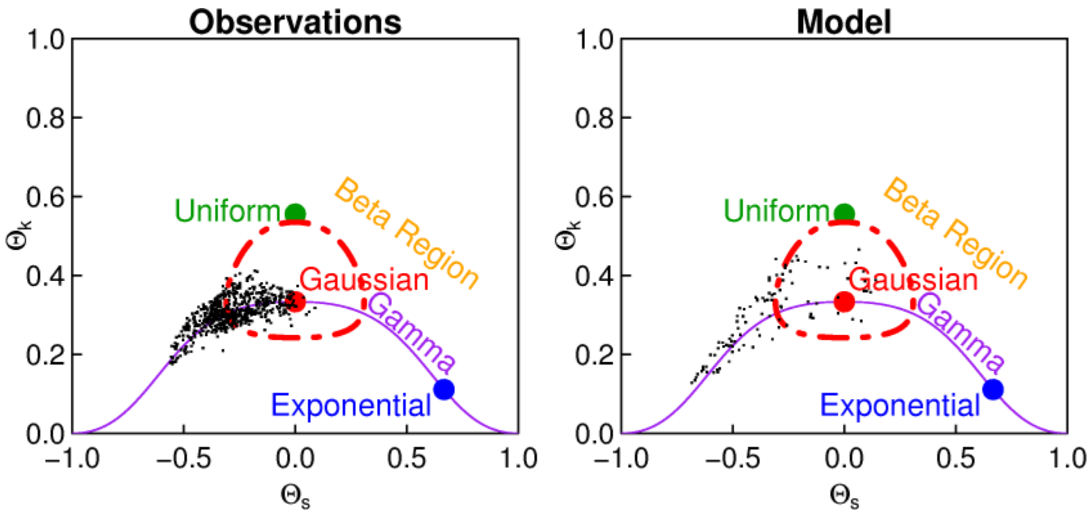

Figure 4

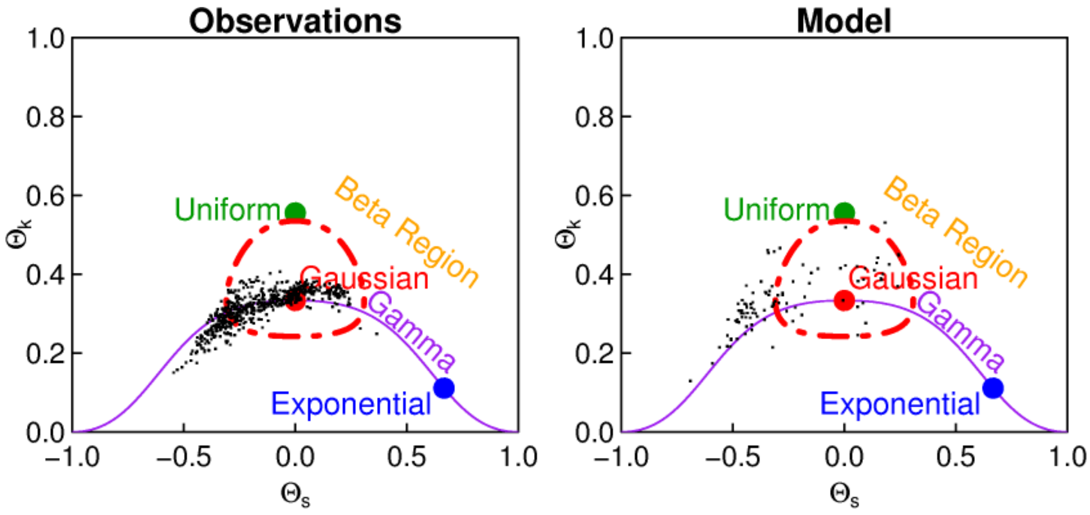

**Figure 5**

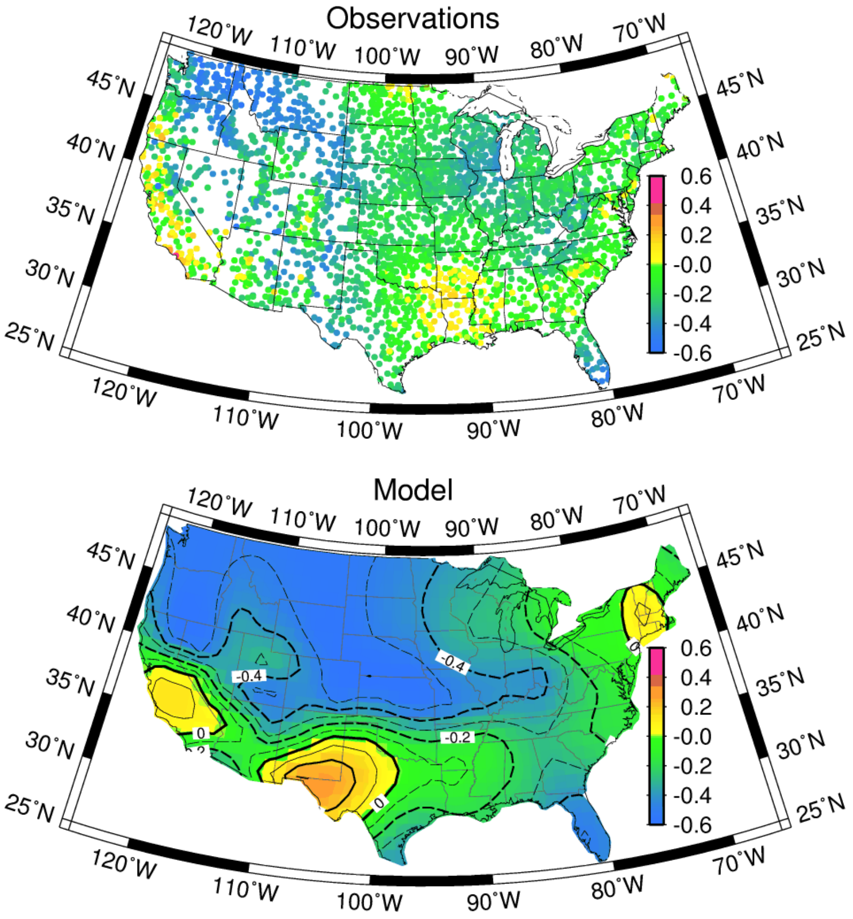
Observations
120°W
110°W
100°W
90°W
80°W
70°W
45°N
40°N
35°N
30°N
25°N
0.6
0.4
0.2
-0.0
-0.2
-0.4
-0.6
Model
-0.4
-0.4
-0.2
0

Figure 6

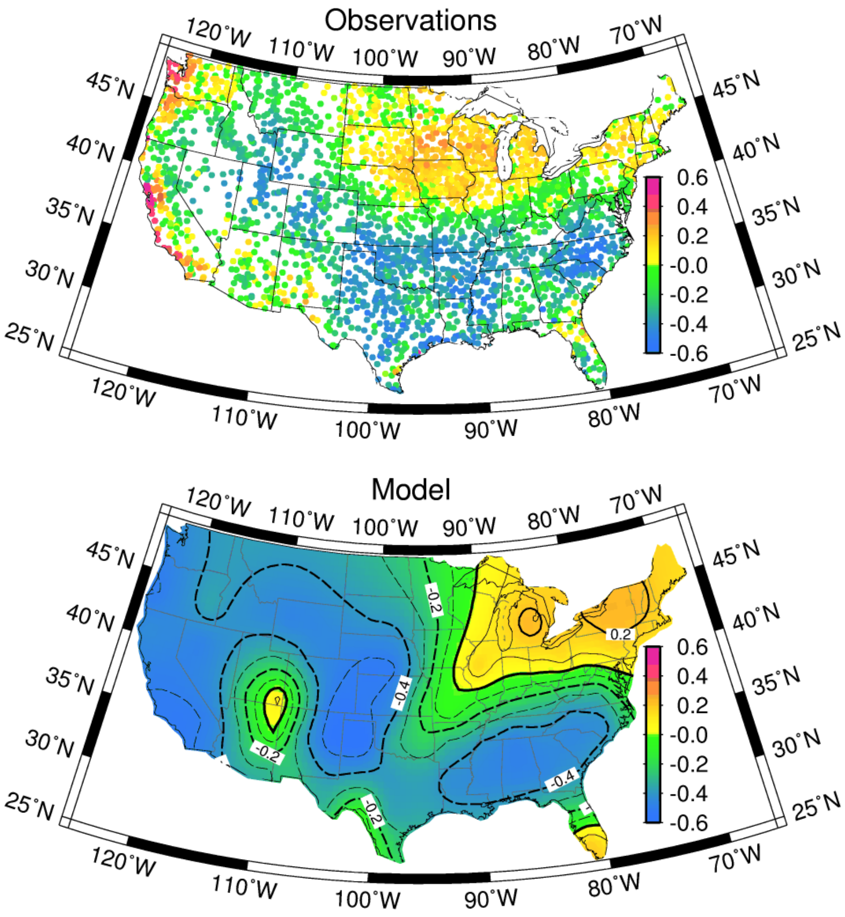
Observations
120˚W
110˚W
100˚W
90˚W
80˚W
70˚W
45˚N
40˚N
35˚N
30˚N
25˚N
0.6
0.4
0.2
-0.0
-0.2
-0.4
-0.6
Model
-0.2
-0.4
0.2
-0.2
-0.4
-0.2

**Figure 7**