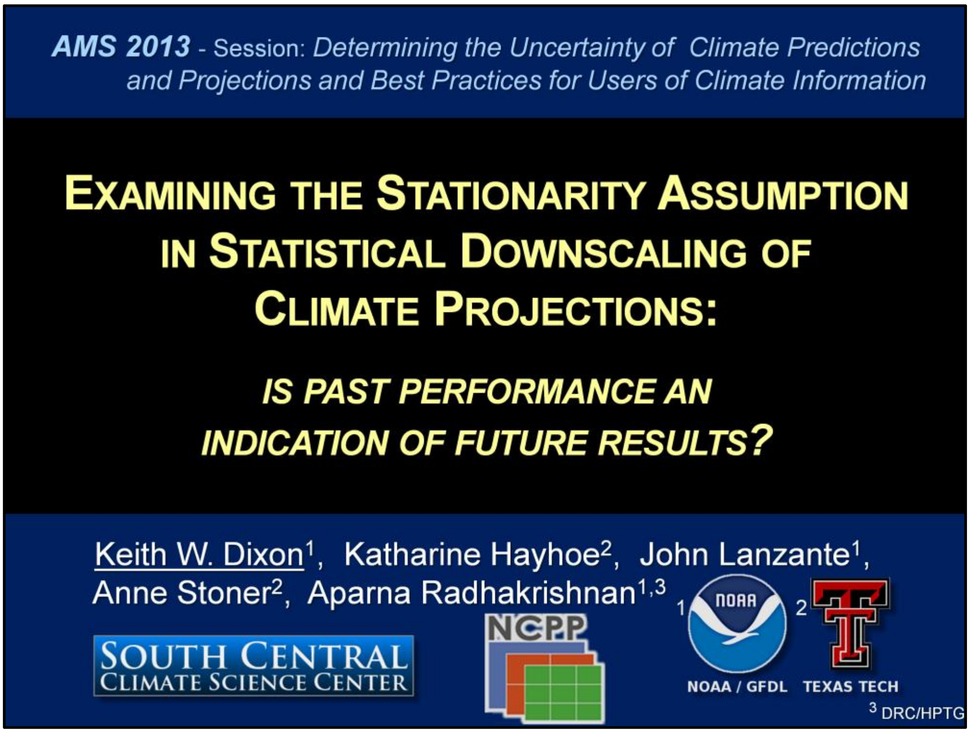
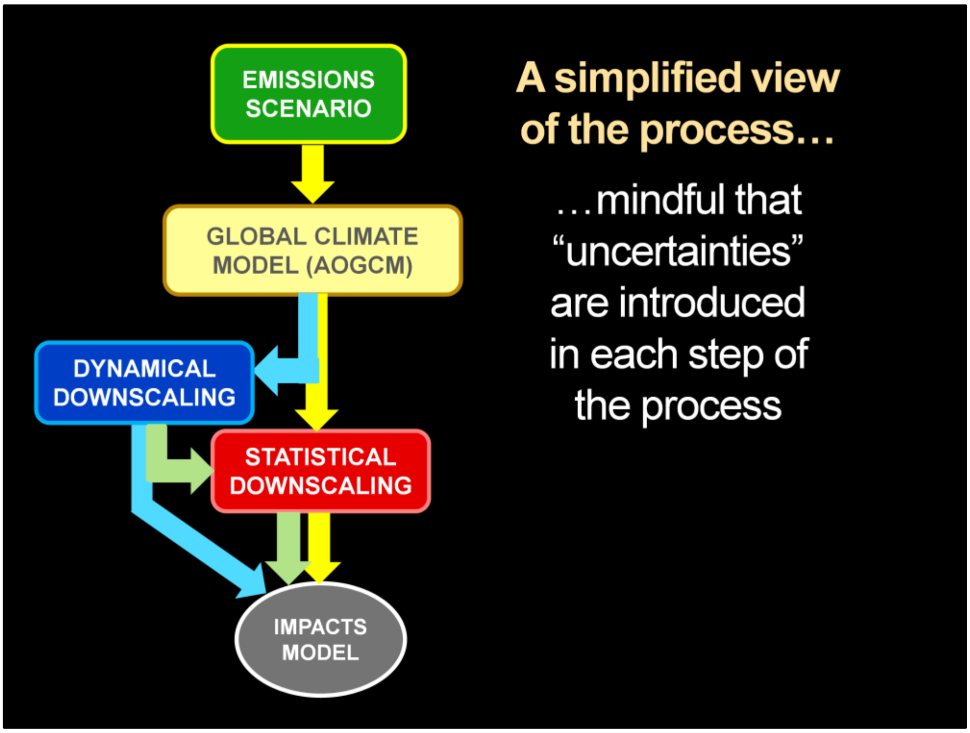


Examining the Stationarity Assumption in Statistical Downscaling of Climate Projections:
Is past performance an indication of future results?
Presented at the 2013 annual meeting of the American Meteorological Society – Austin, Texas – 9 Jan 2013
Abstract: https://ams.confex.com/ams/93Annual/webprogram/Paper221738.html

[slides are those shown at AMS 2013 - notes are expanded beyond AMS presentation]

This presentation summarizes aspects of ongoing work being conducted jointly at NOAA's Geophysical Fluid Dynamics Laboratory in Princeton, New Jersey -and- at Texas Tech University.
Consistent with the theme of this session, we present here a method by which one can examine, in a quantitative manner, one often underappreciated source of uncertainty in statistically downscaled climate change projections – uncertainty arising from what can be referred to as 'the stationarity assumption'.
The research work is funded in part by the US Dept of Interior's South Central Climate Science Center: http://southcentralclimate.org/ &
https://nccwsc.usgs.gov/display-project/4f8c652fe4b0546c0c397b4a/5012e146e4b05140039e03cb
The experimental design protocol and selected results from this project are expected to be part of a National Climate Predictions & Projections (NCPP) Downscaling Workshop in Aug 2013. http://www.earthsystemcog.org/projects/downscaling-2013.
NOAA/GFDL received FY12&13 funds from NOAA-CPO for IT infrastructure development activities associated with this project.

- - -
My email: Keith.Dixon@noaa.gov
My web page: http://www.gfdl.noaa.gov/kd
NOAA's Geophysical Fluid Dynamics Laboratory (GFDL)



The process by which multi-decadal global climate change projections are generated and processed so that they may be used as input to regional or local-scale adaptation planning and climate change impacts studies contains multiple steps, each which has its own set of assumptions, imperfections, and hence uncertainties.

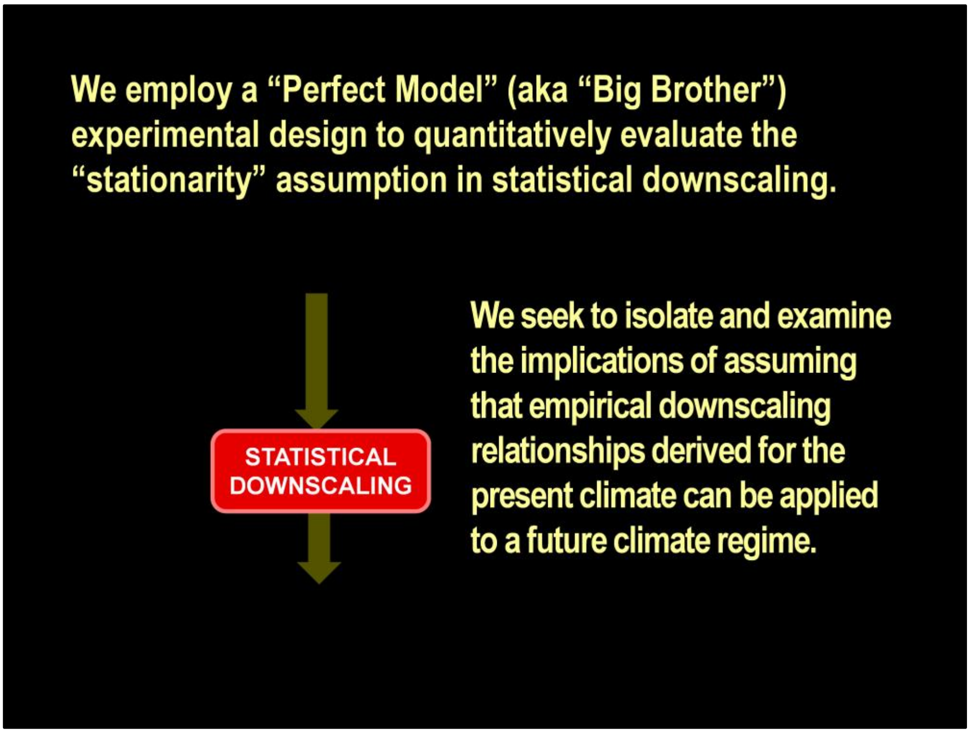
As illustrated here, uncertainties exist regarding the amounts and types of greenhouse gases and radiatively active aerosols that humans will emit in the future. This source of uncertainty is often acknowledged by users of climate information by examining climate model-based projections forced by different emissions scenarios.

Uncertainties in how the global climate system will respond to a given emission scenario is evident in that different global climate models (GCMs) exhibit different climate sensitivities and patterns of response. Users often explore this type of uncertainty by examining output derived from GCMs developed at different research institutions. Projections of future climate generated by different GCMs driven by different emissions scenarios are available to users as part of efforts such as the CMIP5 Data Portal.

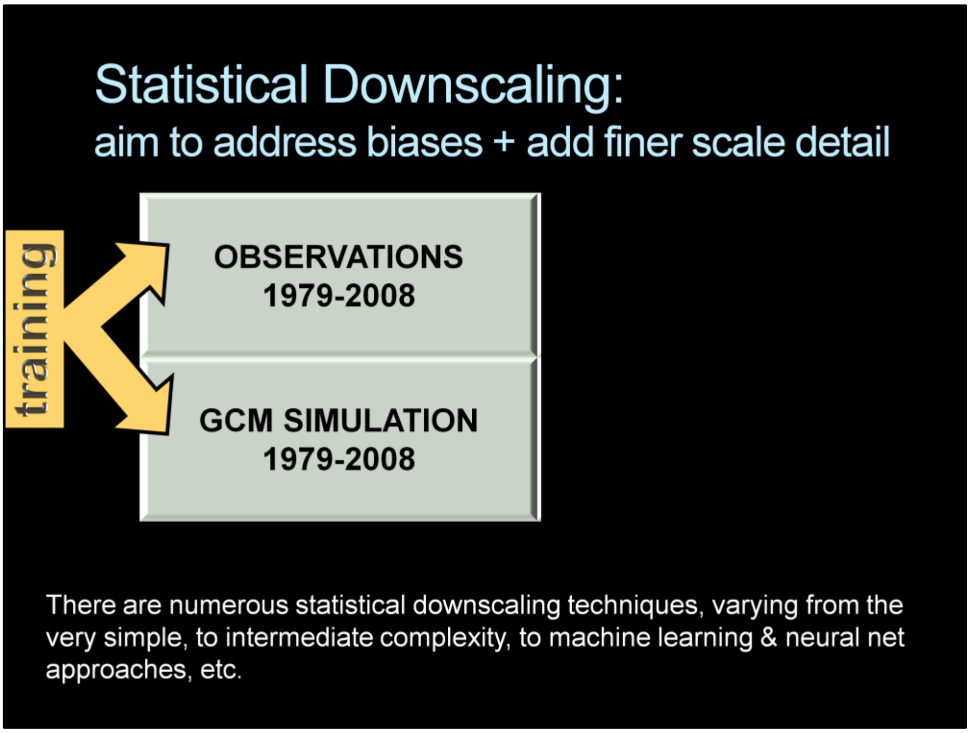
Yet, one way a GCM's results can fall short of the expectations of end users is that GCM output is often on a too coarse grid to provide the desired level of local detail. For example, a GCM with 200km grid spacing can have a single grid cell that overlies ocean waters, beaches, mountains, and everything in between, though the user may be interested in a single location within that area. Additionally, GCM simulations of the contemporary climate may exhibit significant biases at the location of interest (*e.g.,* too hot, too cold, too wet, too dry). Statistical downscaling is often employed in an attempt to account for model biases and to provide additional spatial detail. (For those familiar with operational weather forecasting, you can think of statistical downscaling of climate model output as being a statistical refinement step analogous to the MOS - Model Output Statistics.) A statistical downscaling step, generally considered a "value-added" process, does however, introduce its own uncertainties.

I will also note that statistical downscaling is not the only approach one can take when seeking to generate climate model projections at finer spatial resolution than a GCM. Dynamical Downscaling – the use of finer scale three-dimensional regional climate models driven by GCM output – is another approach being discussed at this meeting, but not in my talk. For some applications, the output of regional climate models is refined further by statistical downscaling.

Related Links:
Downscaling classifications overview: http://en.wikipedia.org/wiki/Downscaling
CMIP5 Data Portal: http://cmip-pcmdi.llnl.gov/cmip5/data_portal.html
MOS: http://www.nws.noaa.gov/tdl/synop/presentations/maryland.ppt

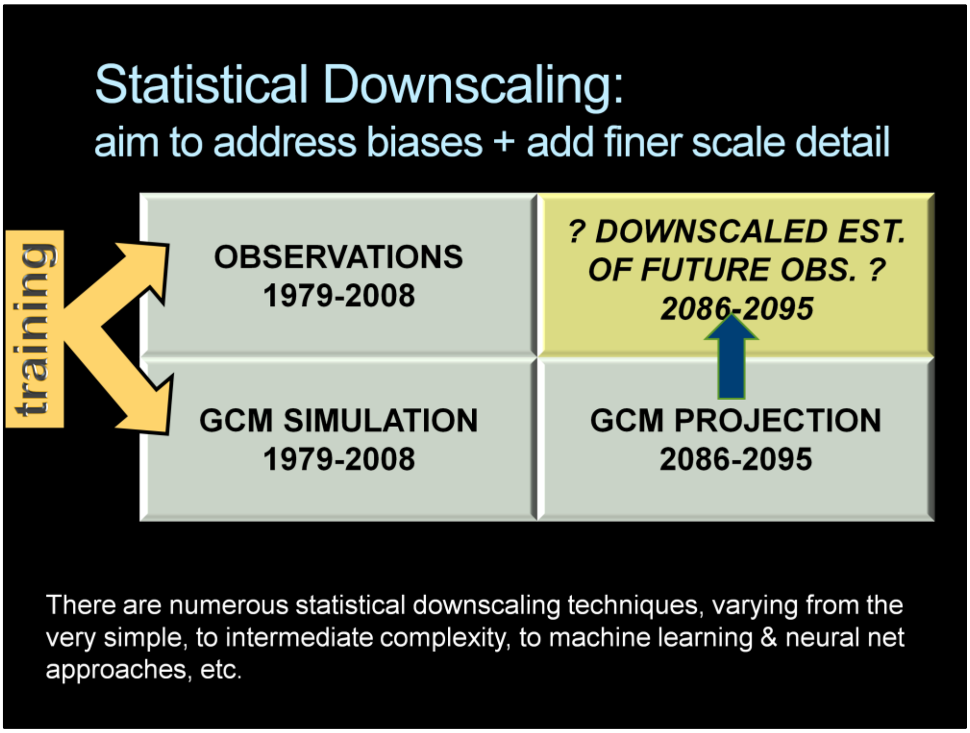


In this talk, we present an experimental design developed to isolate and quantify one particular source (but not the only source) of uncertainty arising from statistical downscaling of climate projections – the stationarity assumption. The stationarity assumption is inherent to statistical downscaling of multi-decadal climate change projections, with the assumption being that statistical relationship between global climate model simulation outputs and real, observed climate data remain constant over time. In other words, we seek to test the assumption that the statistical relationships (aka transform functions) 'learned' during the statistical downscaling training step, are valid when used to generate downscaled estimates of future climate projections. We evaluate the extent to which the stationarity assumption holds by using a 'perfect model' (or 'big brother') framework, as described on the following slides.



Here we illustrate schematically the way in which our 'perfect model' framework isolates and quantifies the extent to which the stationarity assumption holds. We will do so by first describing one typical way that statistical downscaling is implemented in practice. We will contrast that with our perfect model framework, in subsequent slides.

In our example representing a typical real-world application, we start with two data sets - one data set containing observations at the locations of interest for the period 1979-2008 and a companion dataset of GCM model results for the same time period. During the 'training step', statistical methods are used to compare the modeled and observed data sets, and in the process downscaling transform functions are derived that express relationships between the GCM output and the observations. Cross-validation tests performed using the GCM's output (the predictors) and observational data sets (the predictands) from the 1979-2008 era allow one to quantitatively assess how well the statistical downscaling method's transform functions account for GCM biases and finer scale details for the time period for which there are observations.



Having derived transform functions from 1979-2008 observational and GCM data sets during the training step, we can turn our attention to GCM projections of future conditions.

GCM projections of future climate – listed in the example here as being for the 2086-2095 time period – lack finer scale detail one might presume contain biases not dissimilar to those found in the GCM's 1979-2008 simulation.

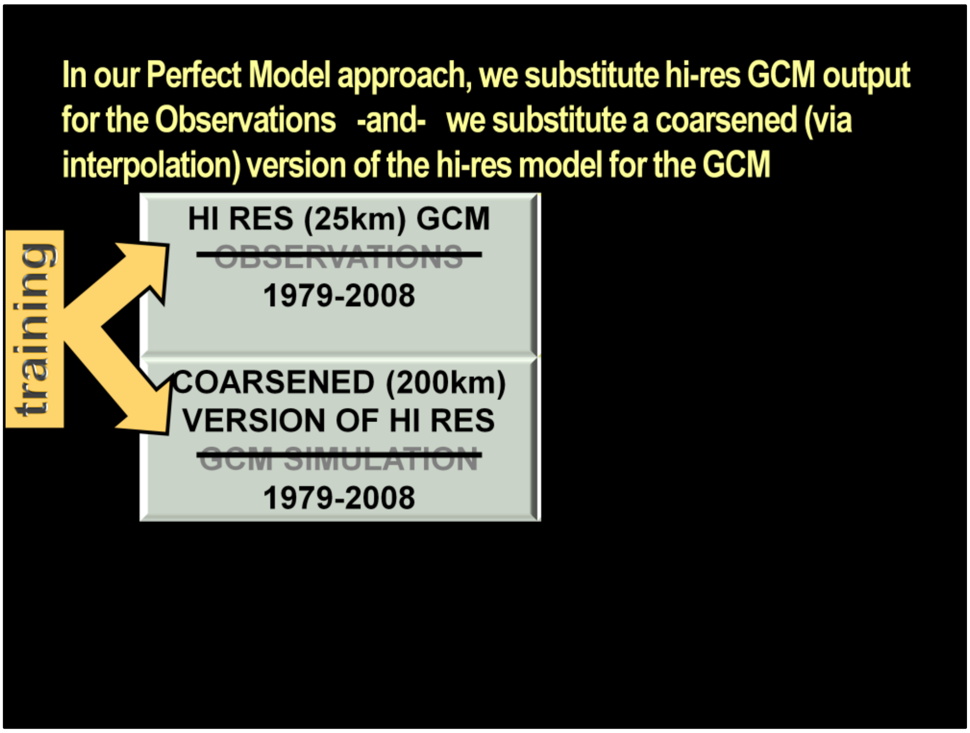
So, one can 'plug' the late 21st century GCM results (used as predictors) into the downscaling equations (transform functions) established during the earlier training step, in order to generate downscaled versions of the late 21st Century GCM projections. Statistically downscaled 21st Century projections generated in this manner can be considered a proxy for future observations in user applications.

Inherent in this process is the assumption that the transform functions derived for the 1979-2008 period apply equally well to the 2086-2095 period. Stated another way, one assumes that, even in the face of a changing climate, what was 'learned' during the training step is applicable to future climate states. This is what we refer to as the stationarity assumption.

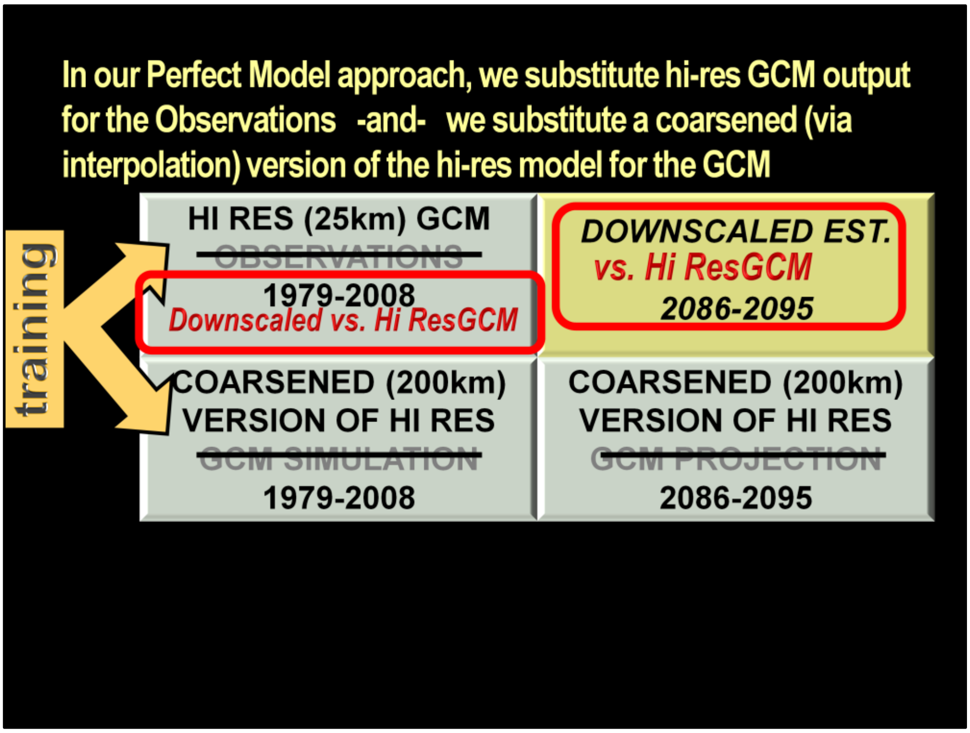
Though, as mentioned above, the skill of a statistical downscaling method can be quantitatively assessed for the 1979-2008 period via cross-validation, the lack of future observations precludes one from conducting a similar downscaling skill assessment for late 21st Century projections. Hence, a direct assessment of the extent to which the stationarity assumption is valid can not be performed using the observational and GCM data sets in real-world applications, as outlined on this slide.

By utilizing different data sets, the perfect model experimental design described on the following slides provides a framework to overcome this diagnostic hurdle.

NOTE: Numerous statistical downscaling techniques exist. They span a wide range of complexity and can differ somewhat in the combination of data sets used in the training step. The schematic outline provided here represents many, but not all, statistical downscaling techniques. However, what they all have in common when used to produce statistically downscaled estimates of future climate projections is that they will employ data sets from the three gray boxes shown in the schematic to generate estimates for the upper right box – and in so doing, they inherently assume stationarity of statistical relationships developed from data drawn from some the gray boxes.

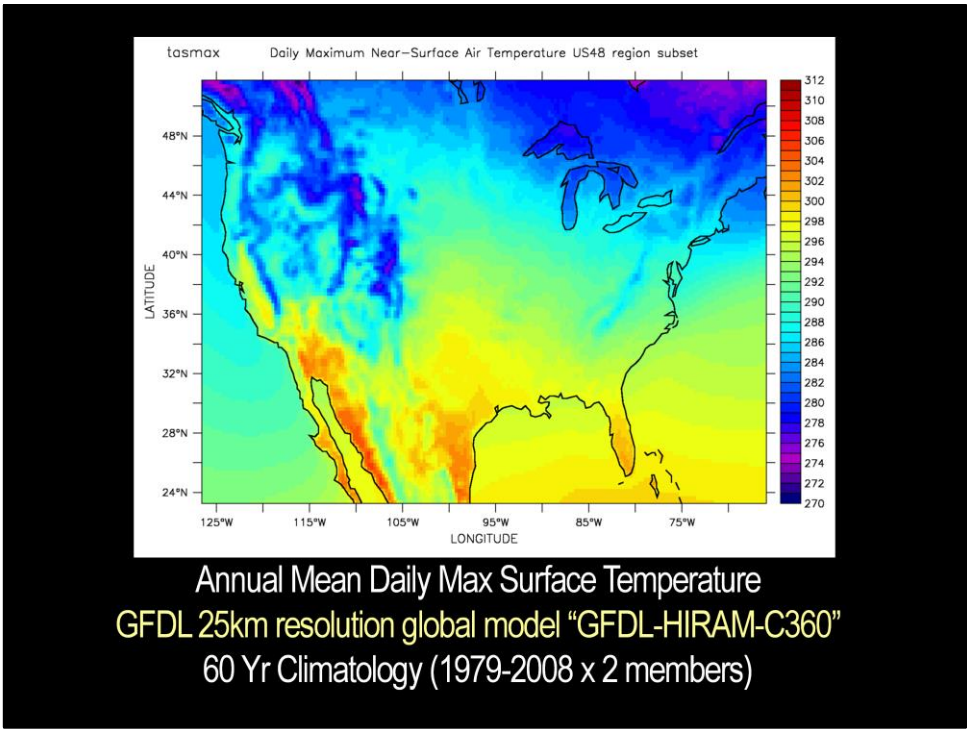


The choice of data sets used is what distinguishes our 'perfect model' experiment design from more typical 'real-world' statistical downscaling procedures, such as the one outlined on the previous 2 slides.
As shown in this revised schematic, in the perfect model framework we substitute high resolution GCM output for observations. (One can think of the high resolution GCM output from a model with ~25km grid spacing as being a proxy for a gridded observational product with similar spatial resolution.) Additionally, we take the high resolution GCM output and interpolate it to a much coarser grid having ~200km grid spacing. In the perfect model framework, this 'coarsened' version of the high resolution GCM data set will take the place of the data sets labeled 'GCM simulation' in previous slides. By using a conservative interpolation scheme, we assure that, when computed over the full spatial domain, area average values of climate variables of interest are identical for the original high resolution GCM output and the coarsened (low resolution) version produced via interpolation. So, while comparison of a high resolution GCM data set and its coarsened by interpolation companion will not exhibit large-scale spatial biases, finer scale details that are lost when interpolating to the coarser grid do result in biases at the grid point level.
In our framework, the statistical downscaling training step that uses as input the 1979-2008 high resolution GCM data set as predictands and a coarsened version of itself as predictors yields transform functions that, in effect, attempt to recover the finer scale details and information lost as a byproduct of the interpolation process.



Moving to the 21st century time frame, one can appreciate an important way in which our 'perfect model' experimental design differs from the usual real-world application of statistical downscaling climate change projections. Unlike the real-world case - for which we lack observations of the future – in this framework we can directly and quantitatively evaluate the skill of the statistical downscaling when applied to future projections by comparing the statistically downscaled results for 2086-2095 to the actual results of the high resolution GCM projections for that time period, grid point by grid point on the 25km grid [indicated by the red box in the upper right quadrant of the schematic]. The stationarity assumption can be said to be perfectly valid if the skill exhibited during the late 21st century period equals that calculated via cross validation of the 1979-2008 period results [indicated by the red box in the upper left quadrant]. Any degradation in skill (increase in the error computed by differencing the high resolution GCM 'truth' and estimates produced by the statistical downscaling) provides quantitative information indicating the extent to which the stationarity assumption does not perfectly hold.

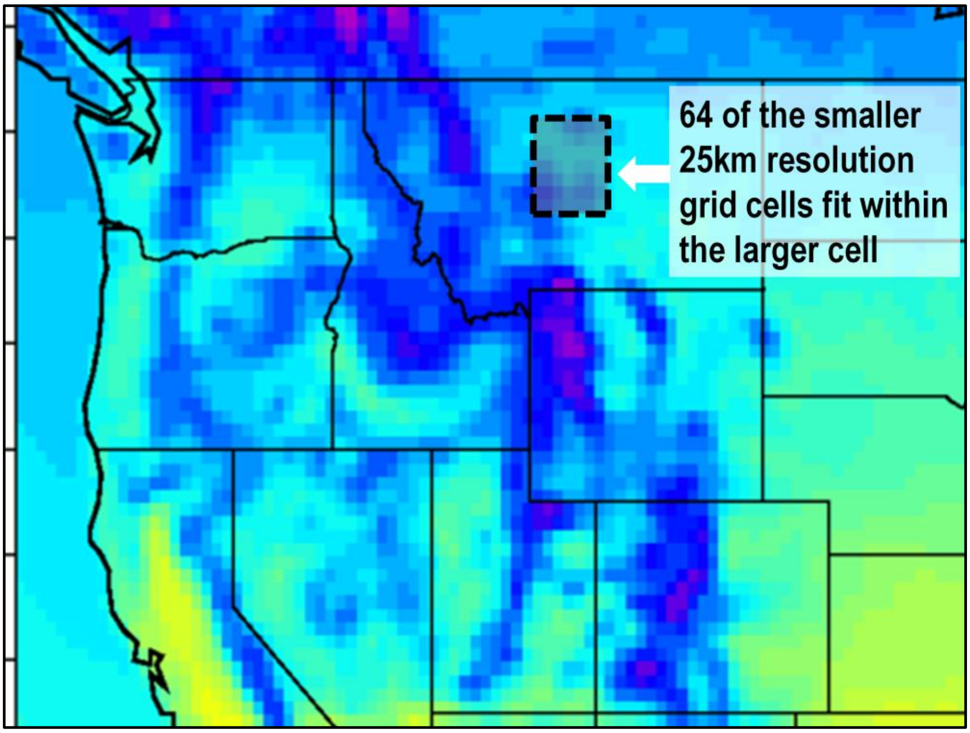
NOTE: The use of the term 'perfect model' should not be mistaken as a claim that a perfect, error-free model exists (neither the GCM nor the statistical model). Rather, it is a term used to identify a type of experimental design or analysis strategy that eschews observational data for model results so as to isolate particular factors and/or to allow assessments of models and methods when observations do not exist. In this case, our perfect model approach uses no observational data. The only data sets required are high resolution GCM results, produced by the same GCM, representing two time periods with different climate characteristics [the upper two boxes in the schematic]. The coarsened datasets [the lower two boxes] are generated simply by interpolating the GCM results to a lower resolution grid (one might consider interpolation to be a very simple model). In turn, the statistical downscaling techniques used to generate high resolution estimates from coarse resolution inputs are statistical models, as well. For evaluation purposes, the late 21st century climate projections produced by the high resolution GCM serve as the 'truth' in this perfect model framework – something not available when using real-world data.



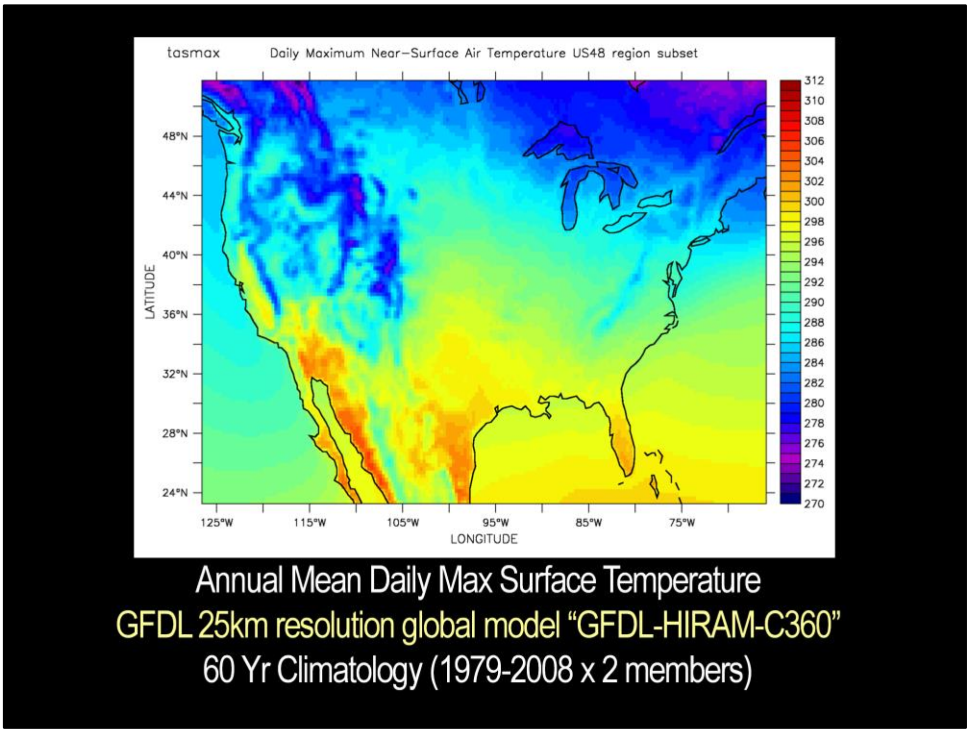
Having outlined our 'perfect model' (or 'big brother') experimental design, we now turn to illustrating how this approach can be applied to garner information about how well the stationarity assumption holds in one particular case. Given time constraints, the analysis presented herein is rather limited, but we hope it provides a sufficient illustration of the technique so that one might better appreciate the method and perhaps prompt some offline discussions.

The high resolution global climate model we use in this example is the GFDL-HIRAM-360 model, which has ~25km grid spacing. The model covers the entire globe, but for illustration purposes we present results focusing on the contiguous 48 United States. Here we see an annual mean climatology of daily maximum temperature calculated from a two-member ensemble of AMIP-style experiments run for the time period 1979-2008 (60 years total).

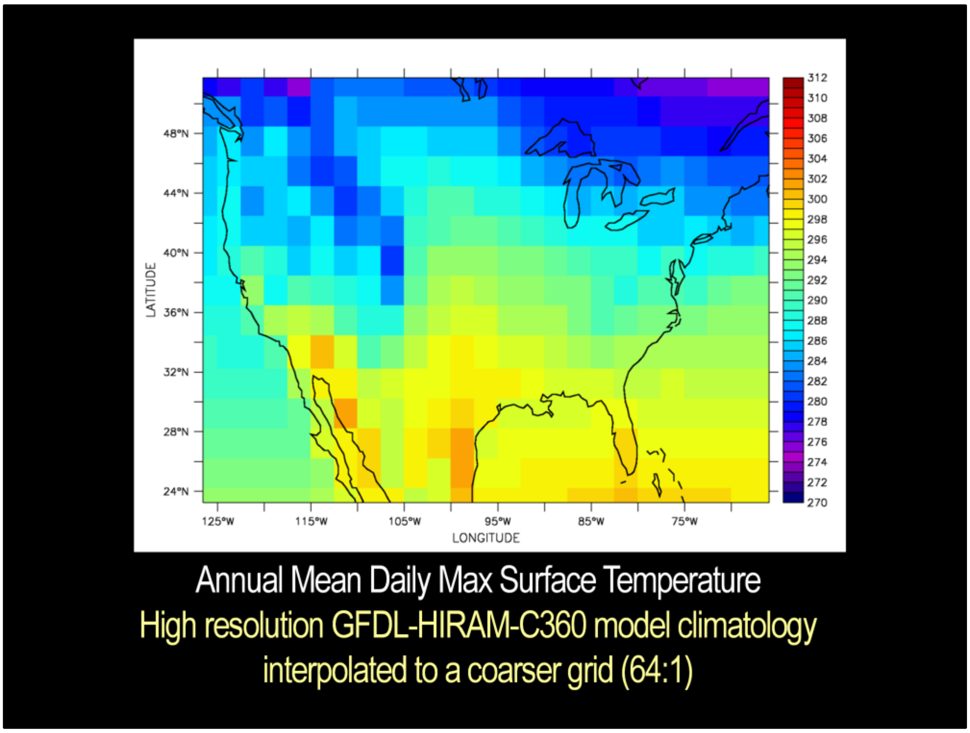
Note: For completeness, we note that the GFDL-HIRAM-C360 experiments shown here were run as AMIP-style experiments following CMIP5 protocols. This means that rather than having the GCM's atmosphere and land surface components coupled to fully dynamical ocean and sea ice model components, observed time-varying sea surface temperatures (SSTs) and sea ice values were specified as lower boundary conditions. That the climate model we use to illustrate the perfect model experimental design was run in this manner, rather than as a fully coupled dynamical atmosphere-ocean model, does not alter the qualitative results presented here. The perfect model framework applies just as well to fully coupled GCMs.



Here we zoom in on the northwestern US, so that one might visually compare the high resolution ~25km grid of the C360 model output (individual color filled boxes) and the approximate size of a grid cell on the coarser grid (~200km grid spacing).
About 64 of the smaller ~25km grid cells fit within the larger ~200km grid cell.

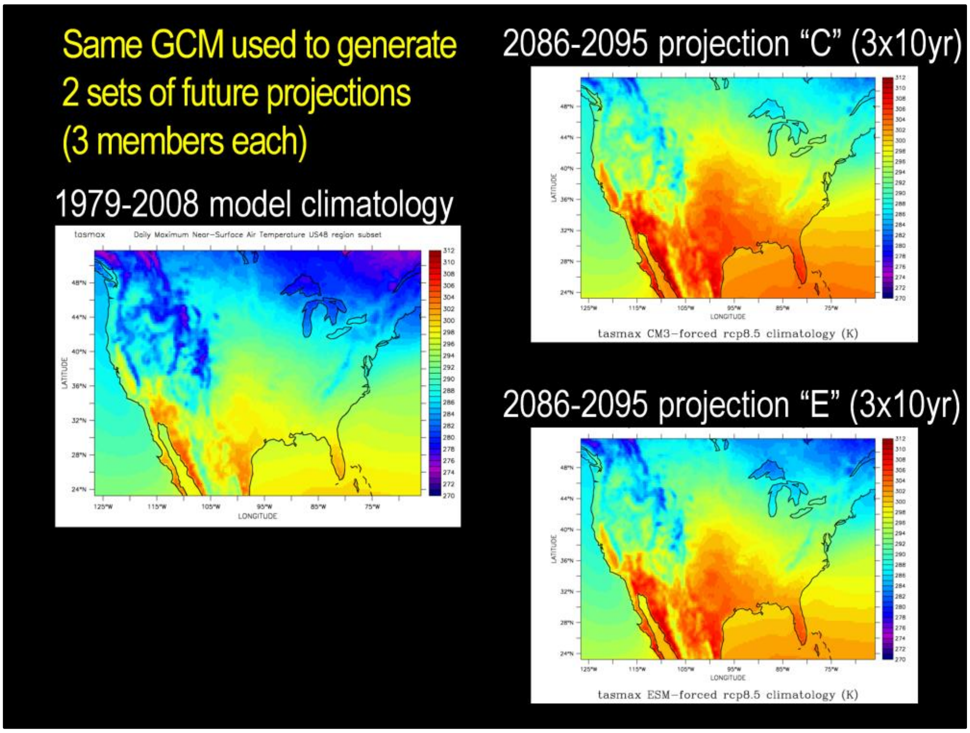


When we apply conservative first-order interpolation, the annual mean model-generated climatology for daily maximum temperature goes from looking like this…

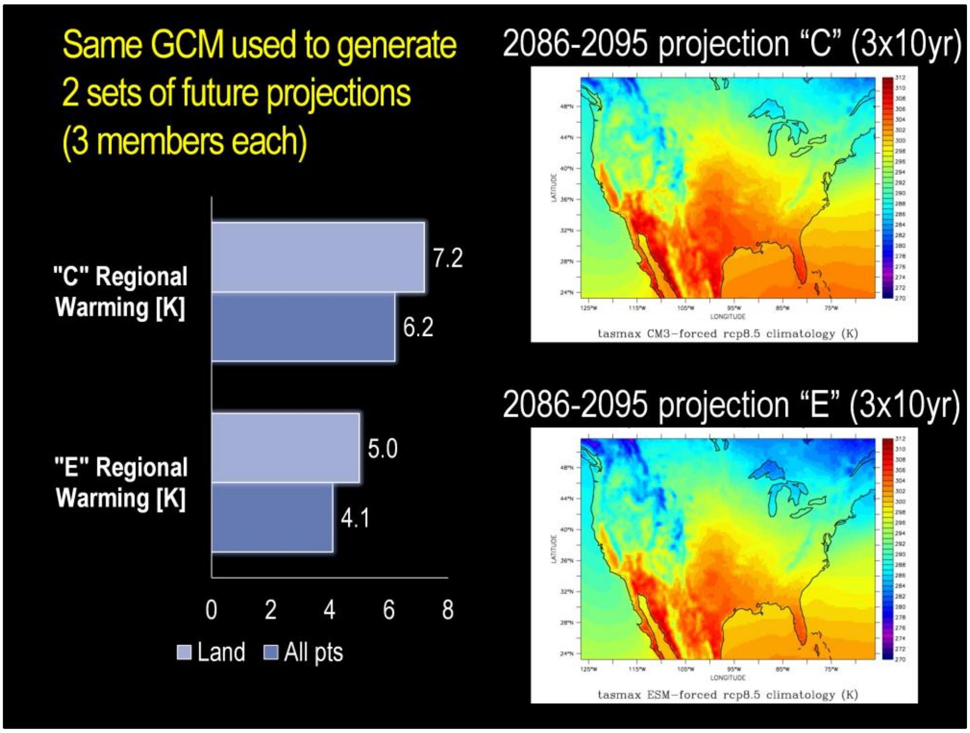


…to this.

The area mean computed across the entire domain is the same for the two data sets having different spatial resolution, however the loss of information incurred due to interpolation is readily evident when comparing this slide with the previous slide.

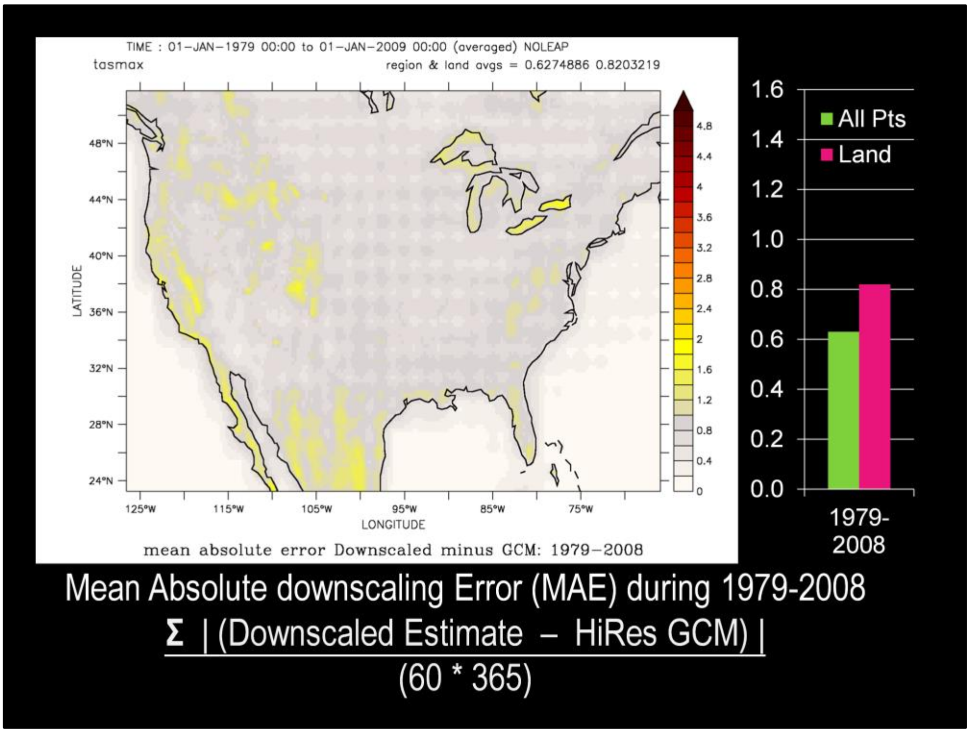


Next, the same GFDL-HIRAM-C360 model code was run in ‘time-slice mode’ to produce sets of future high resolution climate projections. For the purposes of this illustration, suffice it to say that two sets of C360 (high resolution GCM) projections were generated for the period 2086-2095, with each set consisting of three ensemble members. Both sets of projections exhibit a sizable amount of warming by the late 21st century. One set, labeled “C” here, warmed more than the other set labeled “E”.



As shown in the bar charts, the warming of daily maximum temperatures in the set of "C" experiments averaged slightly more than 2 degrees Celsius greater than that simulated in the set of "E" experiments, when averaged over the region mapped. Considering only land points, daily maximum temperatures in the set of "C" experiments on average warmed 7.2C vs. 5.0C of warming in the set of "E" experiments. Daily maximum temperatures over land warmed more than over ocean points, as evident in the "over land only" average being about 1C greater than that computed over the entire geographic region shown in the maps.

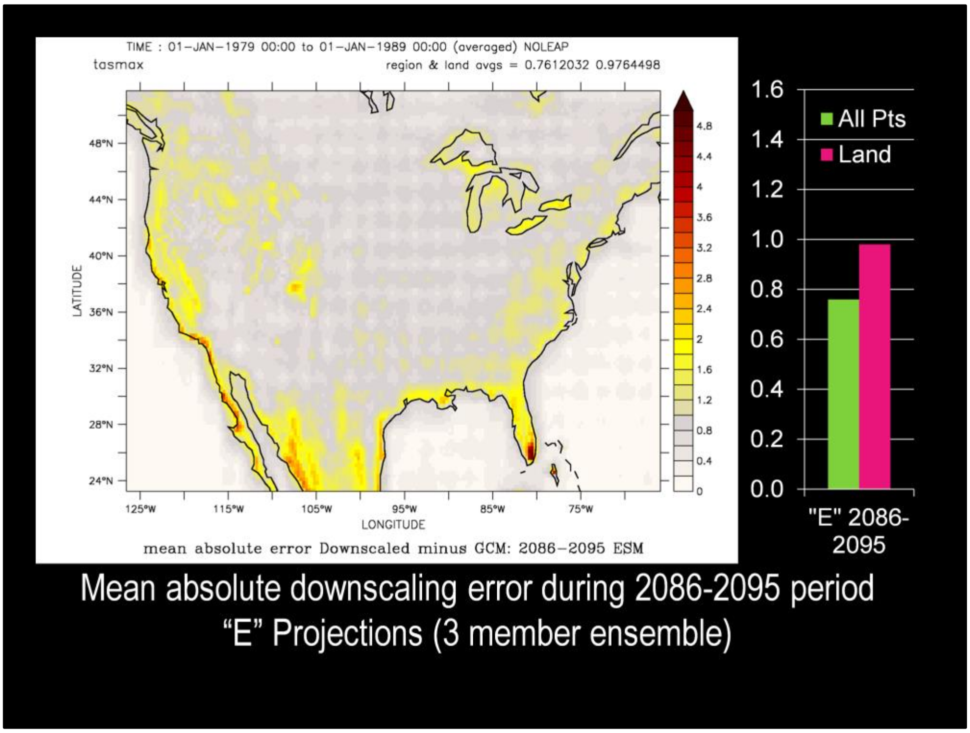
Note: For completeness, we note that for time-slice simulations of the future, projected changes in SSTs and sea ice simulated by fully coupled global GCMs (GFDL CM3 = "C" and GFDL ESM2M = "E") are imposed as lower boundary conditions for the high resolution C360 model.



On this map, we show one measure of how well a particular downscaling method was able to recover the full detail of daily maximum temperatures simulated by the high resolution model during the 1979-2008 period. We refer to the measure shown as the "Mean Absolute downscaling Error", or "MAE". At each high resolution grid point and for each day in the GCM simulations we compute the difference between the actual high resolution GCM's daily maximum temperature and the estimate for that day produced by plugging the coarsened version of the model output into the previously computed downscaling transform functions. We then take the absolute value of that downscaling error for each day and compute the average over all 21,900 days (2 ensemble members x 30 years per ensemble x 365 days per year) and map the results.

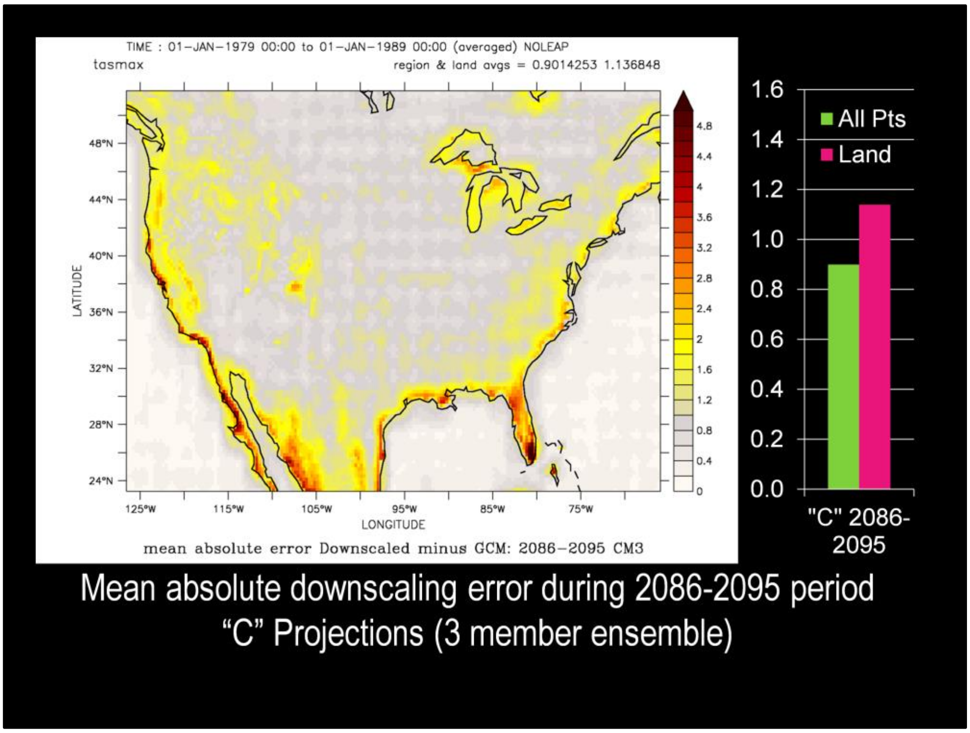
One can see that smaller MAE values tend to occur over the open ocean and that larger MAE values tend to occur along coastlines and areas of step topography – places one might expect to have larger horizontal gradients in the high resolution data that could be poorly represented when interpolated to a much coarser grid.

We can consider this map to serve as a baseline, against which MAE values computed for the 2086-2095 period can be compared. If the stationarity assumption holds, the MAE maps computed for the late 21st century projections would not exhibit significantly larger MAE values.

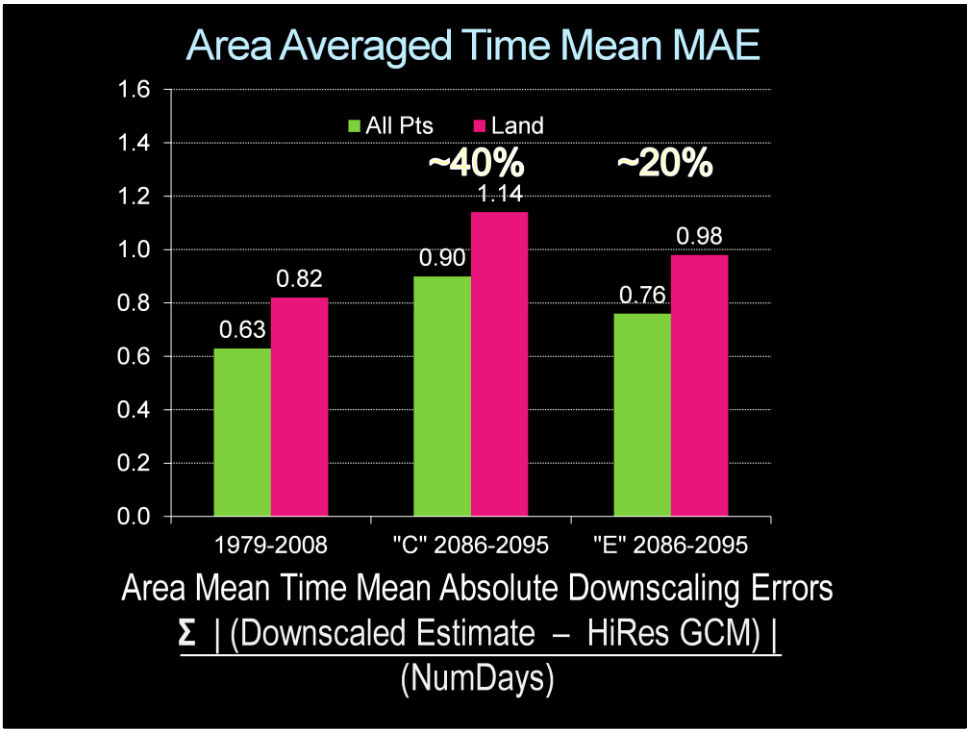


Here we show the time mean MAE field for daily maximum temperature averaged over the 3 members of the “E” ensemble (the set of late 21st century projections that warmed on average 5.0C over land points). Visual inspection suggests that the MAE of this set of late 21st century experiments is greater than that computed for the 1979-2008 period. A similar pattern of smaller errors over the ocean and larger errors along coasts and mountainous regions is seen in this map and the previous map of the 1979-2008 period.

(Recall that the downscaling transform functions were computed by comparing the coarsened predictors to the high resolution GCM’s daily maximum temperatures [predictands] for the 1979-2008 time period and that late 21st century statistically downscaled estimates were produced by plugging coarsened versions of the 2086-2095 GCM fields into the transform functions computed for the earlier period.)



And here we show the time mean MAE field for daily maximum temperature averaged over the 3 members of the "C" ensemble (the set of late 21st century projections that warmed on average 7.2C over land points). Visual inspection suggests that the MAE of this set of late 21st century experiments is greater than those computed for the set of later 21st century "E" experiments and the 1979-2008 period. Some of the larger MAE values are seen over peninsulas (Florida and Baja California).



The bar graphs presented here document the area average MAE values computed over the geographic domain shown in the three previous maps. For each of the three pairs of adjacent bars, the bar on the left represents the average MAE computed over all points (land and ocean) and the bar on the right represents the average MAE computed over land points only.

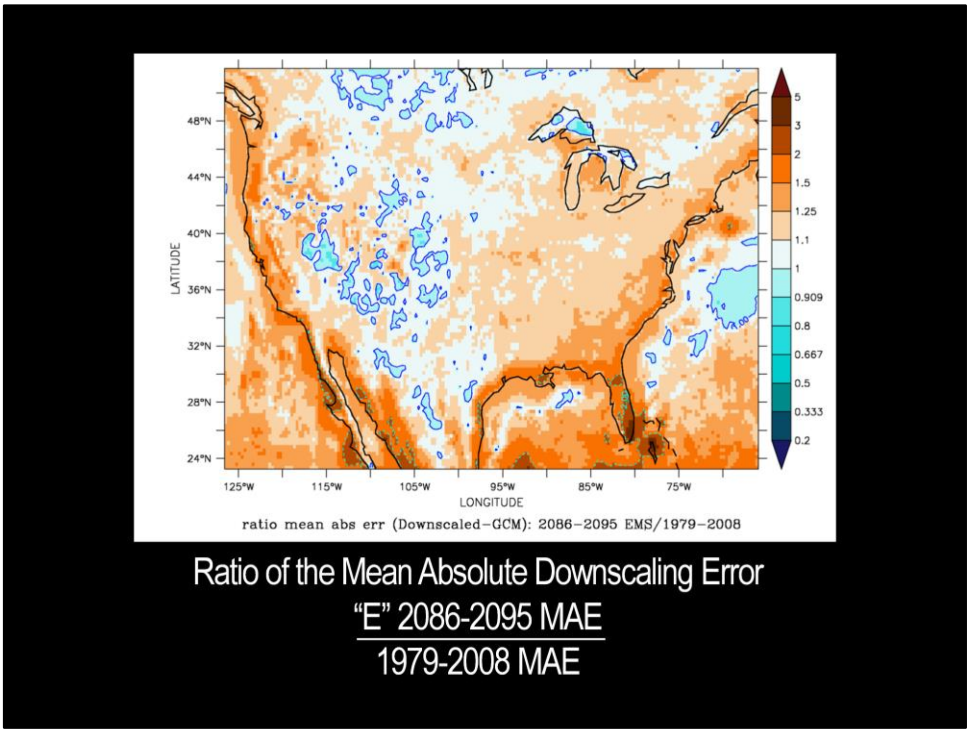
The leftmost of the three pairs of bars shows MAE values for the baseline 1979-2008 case.

Comparing the baseline results (left) with the rightmost pair of bars reveals that average MAE values increased by 20% in the “E” experiments that experienced a mean warming of daily maximum temperature over land of 5.0C.

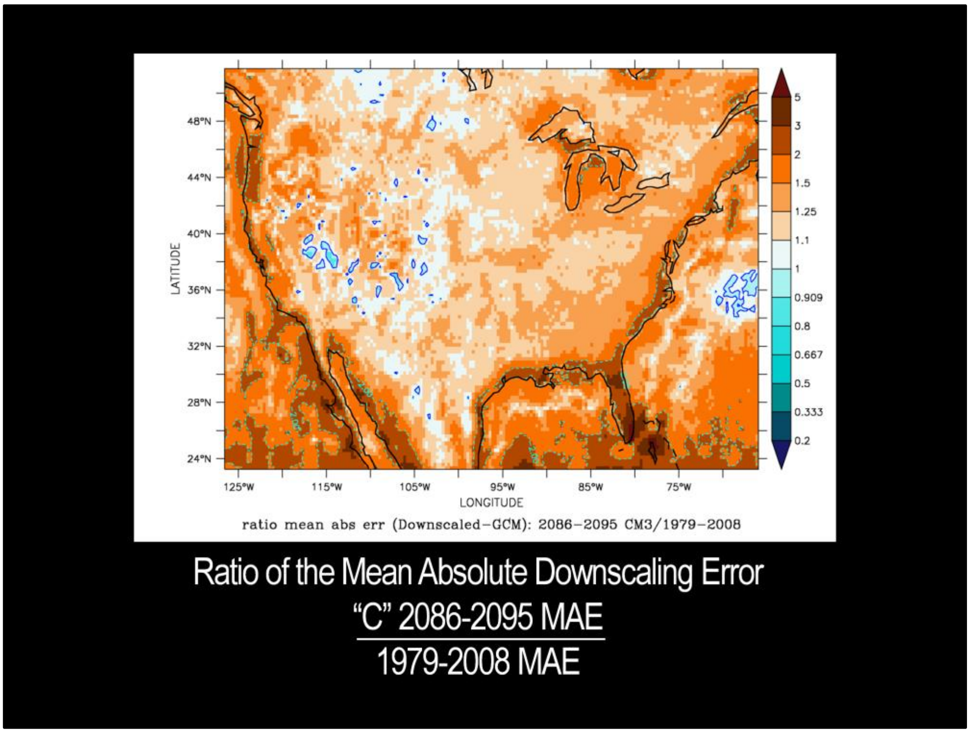
And comparing the baseline results (left) with the center pair of bars reveals that average MAE values were 40% greater in the “C” experiments that experienced a mean warming of daily maximum temperature over land of 7.2C.

The relative increase in the area averaged MAE was similar over land and over ocean points for both the “C” and “E” cases.

From these results, one can conclude the statistical relationships (aka transform functions) ‘learned’ during the statistical downscaling training step used in this particular example produces larger errors when used to generate downscaled estimates of future climate projections than when applied to the same time period used in the training. Thus, one can state that the stationarity assumption is not completely valid, though the amount by which the error could grow and still be considered to produce a “value added product” is subjective. It is not our intent here to render such judgments. Nor are we suggesting that the MAE should be considered the error metric of choice when making such assessments. Rather, our goal is to present an experimental design that allows one to quantitatively assess the issue, and to provide examples of analyses that illustrate the potential of this experimental design.



Next we illustrate geographical differences in the extent to which the MAE error statistic grows when going from the 1979-2008 base period to the late 21st century projections. We do so by computing and mapping the ratio of the MAE computed for the late 21st century period to the MAE computed over the base period.
This map shows the results for the set of "E" experiments (the ones in which daily maximum temperatures warmed ~5.0C on average over land).
The blue contour is drawn for ratios of 1.0, indicating that the few areas shaded in light blue did not experience an increase in MAE for daily maximum temperature. Note the non-linear color bar. Areas having the lightest shading (nearly white) had time mean MAEs grow by less than 10%. Areas shade the light orange had MAE grow from 10 to 25%, etc. A faint dashed contour encompasses areas in which MAE values computed over 2086-2095 were more than double that of the base period.



In general, the non-uniform pattern of relative MAE increases seen in the "C" case shown here (7.2C warming over land) appears to be an amplified version of that seen in the "E" case (previous slide) that experienced ~2C less warming.

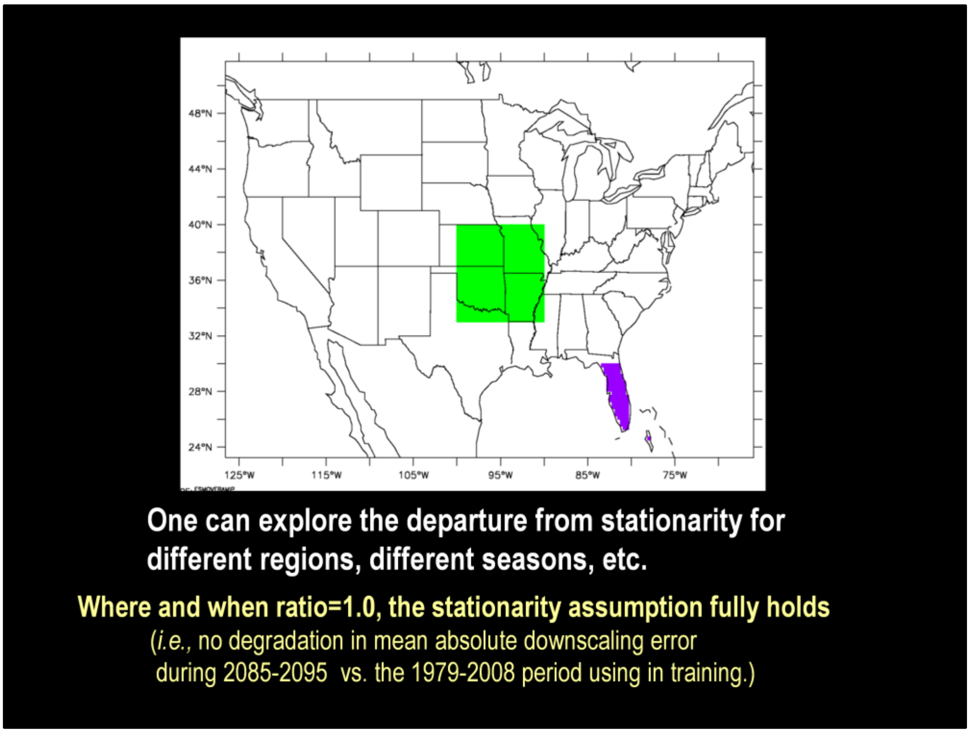
Earlier, we noted that the largest MAE values tended to occur along coastlines and mountainous areas. Interestingly, we see here that the MAE Ratio tends to increase much more along coasts than in interior mountain regions. This observation has prompted us to initiate additional analyses reviewing the nature of the downscaling errors along coasts in the hope of gaining an understanding of whether this behavior arises from a weakness in the particular downscaling method used to generate these results -or- from a change in GCM-simulated climate patterns or atmospheric circulation that only expresses itself in the late 21st century, and thus can not be readily detected by statistical downscaling training that uses 1979-2008 data sets as input.

Note: The statistical downscaling method employed in these examples is an asynchronous regional regression model (ARRM) - a daily quantile regression approach documented in…

Stoner, A. M. K., Hayhoe, K., Yang, X. and Wuebbles, D. J. (2012), An asynchronous regional regression model for statistical downscaling of daily climate variables. Int. J. Climatol.. doi: 10.1002/joc.3603

http://onlinelibrary.wiley.com/doi/10.1002/joc.3603/abstract

See also http://cida.usgs.gov/climate/hayhoe_projections.jsp



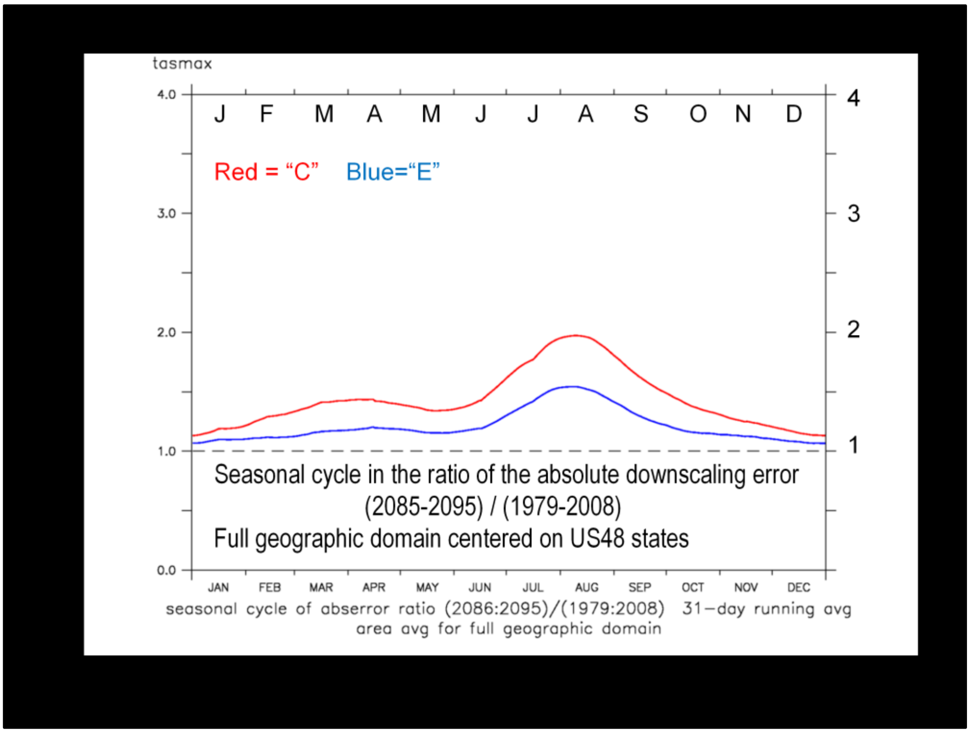
Of course, looking at annual mean error statistics is not the only way one can examine validity of the stationarity assumption.

In the next three slides we show seasonal variations of the MAE Ratio for different geographic regions.

The first will display results averaged over the entire domain shown here.

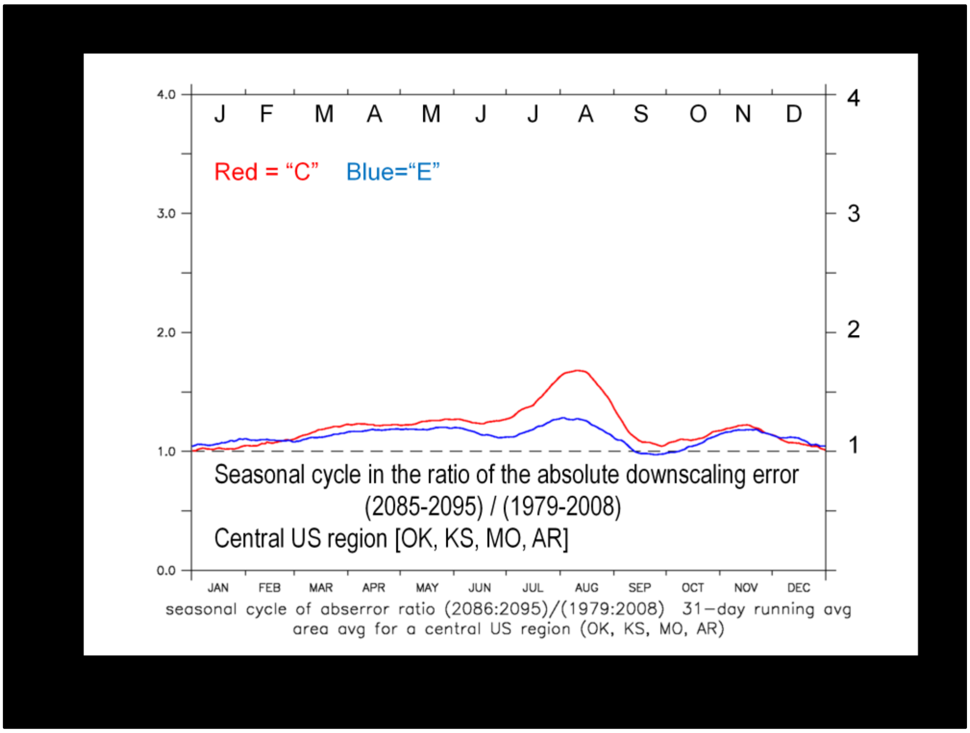
The second will focus on the area average for the central US region shaded green – an area that had relative low annual mean MAE values.

The third will focus on the Florida peninsula area shaded here – an area that had relatively large annual mean MAE values and large MAE ratios.

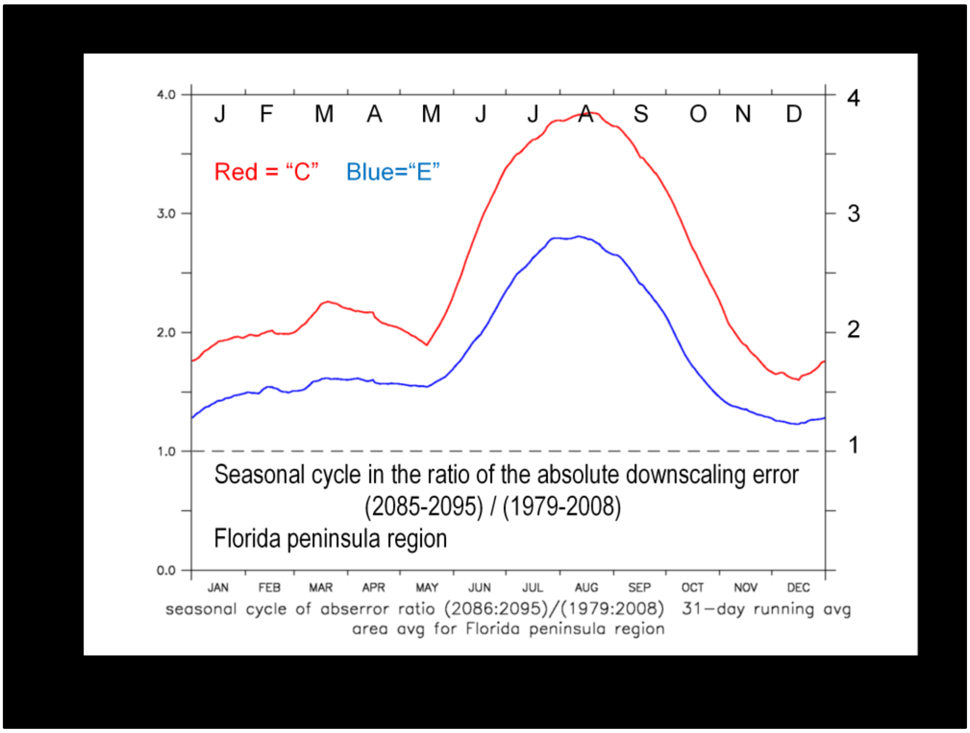


For the full geographic domain, we see that the stationarity assumption holds fairly well during the winter months and that the largest contribution to the annual mean MAE Ratio increase comes from August. We also see that the general seasonal pattern in similar in the “C” (red) and “E” (blue) experiments, though the “C” case exhibits greater increases in the ratio year-round.

Note: In this graph and the two that follow, the seasonally-varying climatologies of MAE ratios were computed day-by-day from the 30 years available for each of the “C” and “E” ensembles of late 21st century projections and from the 60 years available from the two 1979-2008 experiments. A 31 day boxcar smoother was applied to remove shorter-term noise from the plotted curves.



Here we see results for the central US area – a region that was characterized by relatively low MAEs compared to other land areas. It is interesting to note the summertime increase in MAE Ratio is readily apparent in the "C" experiments that experienced more warming, but less so in the "E" experiments.



For the Florida Peninsula region – an area of relatively large downscaling errors – we see that the increase in the errors during the late 21st century relative to those during the 1979-2008 period are much more pronounced in the summer than in the winter. Again we see that the larger the amount of warming, the greater the increase in the downscaling errors ("C" errors are greater that "E" errors).

Examining the Stationarity Assumption in Statistical Downscaling of Climate Projections:
Is Past Performance an Indication of Future Results?

- In the example shown, the skill of the downscaling method was diminished in the late 21st century projections relative to the 1979-2008 period.
- Changes in downscaling skill varied seasonally, geographically, and were greater in the projection exhibiting a more pronounced warming response.
- Initial analysis informs additional looks at physical climate and downscaling methodologies.
- Not shown, but to be explored in the Big Brother framework… how well the stationarity assumption holds for different statistical downscaling techniques, different climate variables, different skill metrics, extreme events, etc.

The bullets on this slide list some of the general findings we have presented. We hope the presentation was successful at conveying information about our perfect model framework and how it can be applied to assess the validity of the stationarity assumption in applications of statistical downscaling -and- that one can appreciate that analyses of this type can also yield information about the characteristics of the downscaling techniques themselves (insights that may feedback to inform development of refined versions of the techniques).

In summary, the intent of this presentation was to…

(a) describe what the ‘stationarity assumption’ is as it relates to creating statistically downscaled climate change projections and why it matters.
(b) outline a ‘perfect model’ (aka ‘big brother’) experimental design that allows one to isolate and quantify the extent to which the stationarity assumption holds for a given application.
(c) begin to illustrate the kinds of analyses one can perform using results of the perfect model framework.

What is presented here only scratches the surface of what can be examined using this experimental design. Questions we will continue to explore include…

(i) How well do different downscaling techniques perform with respect to the stationarity assumption when applied to the same perfect model data sets.
(ii) Examine a wider range of climate model outputs (temperature, precipitation, surface radiation, soil moisture, etc.)
(iii) Apply different measurements of skill to the output of the perfect model experiments, for example, methods that focus on the performance for extreme events.
(iv) One can apply the perfect model experimental design using output from different GCMs, thereby testing whether the statistical downscaling results are sensitive to the underlying GCM.