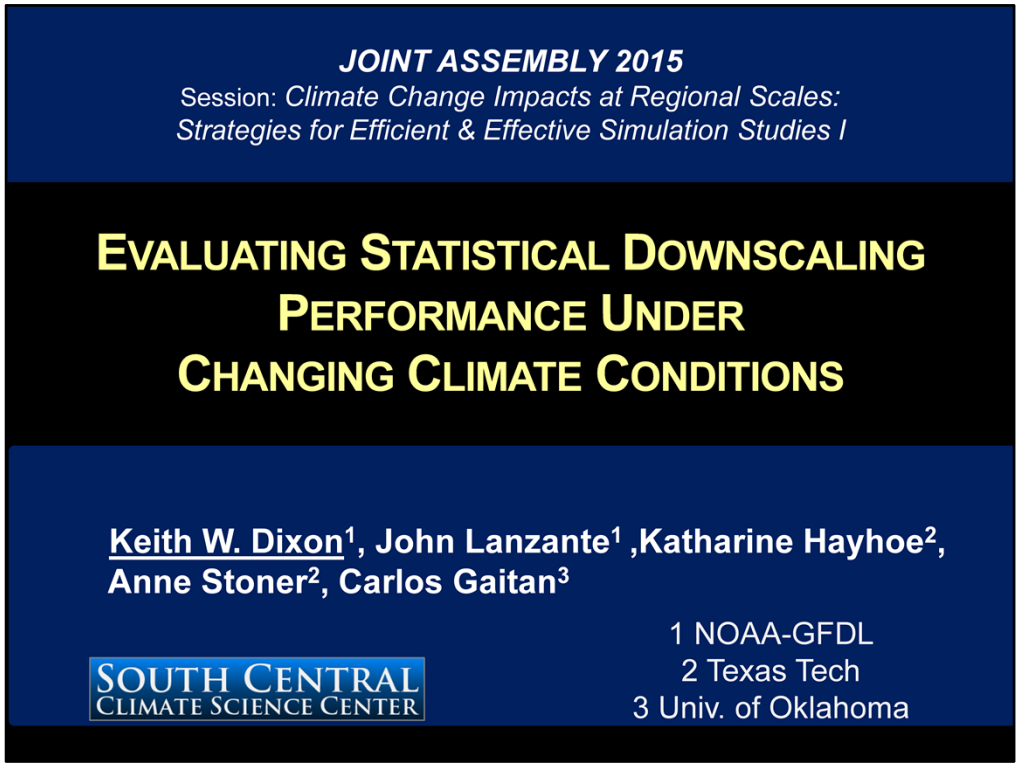
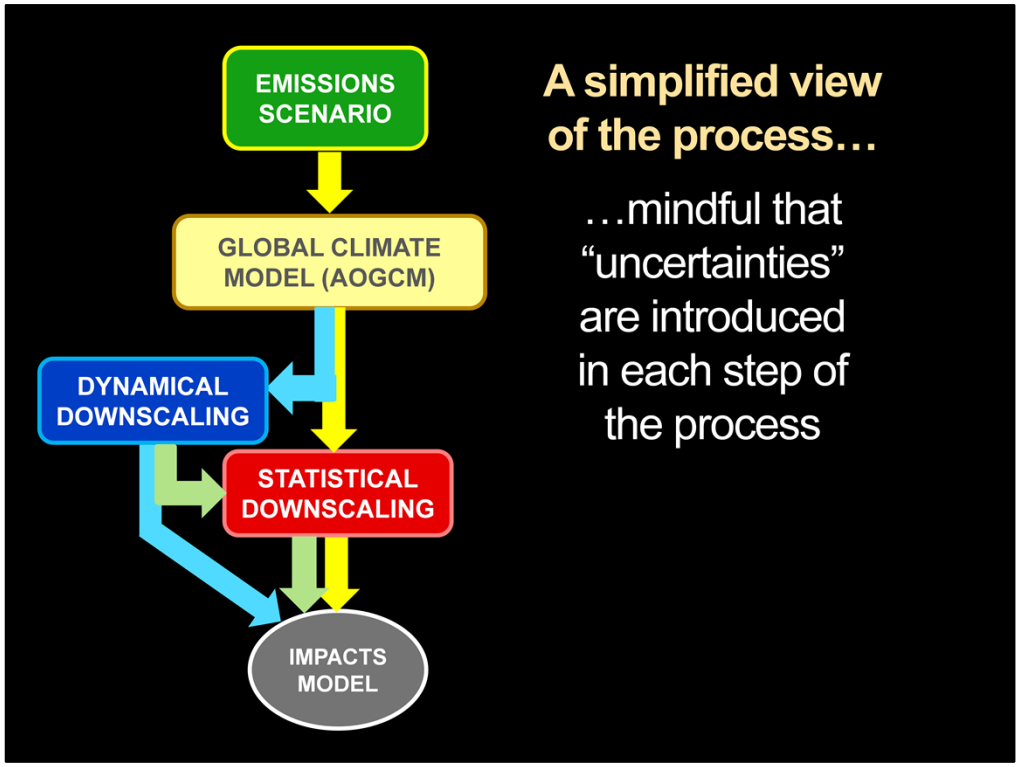


My email: Keith.Dixon@noaa.gov
GFDL's Emprical Statistical Downscaling Team's web page: http://www.gfdl.noaa.gov/esd
NOAA's Geophysical Fluid Dynamics Laboratory (GFDL)



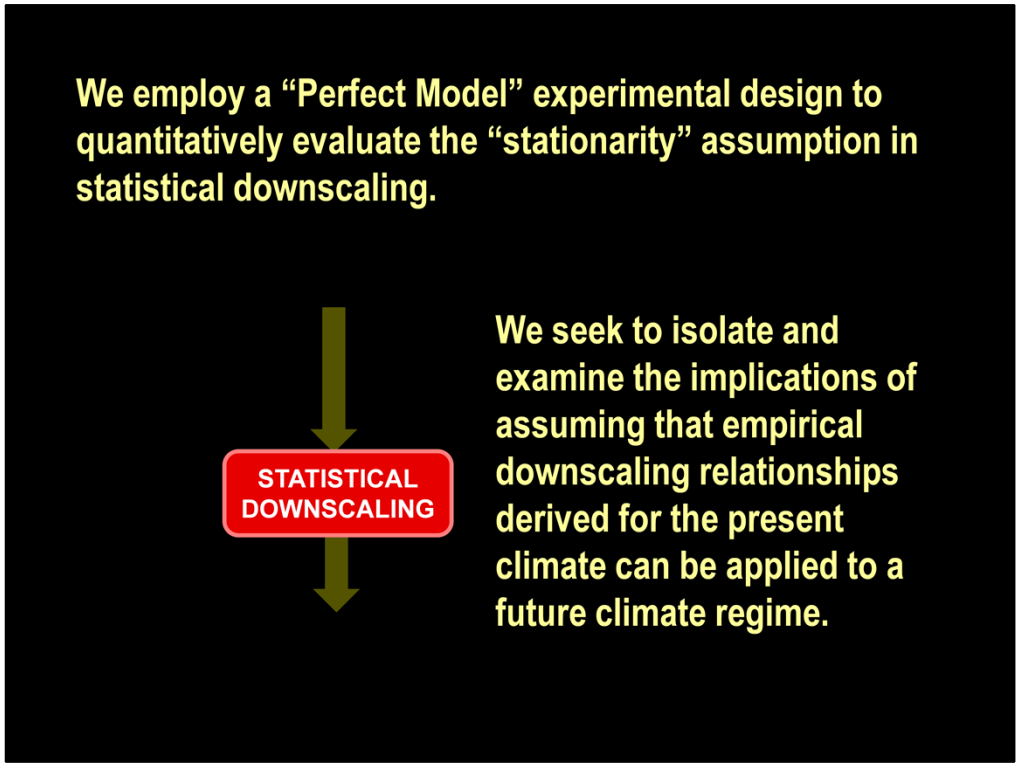
The process by which multi-decadal global climate change projections are generated and processed so that they may be used as input to regional or local-scale adaptation planning and climate change impacts studies contains multiple steps, each which has its own set of assumptions, imperfections, and hence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

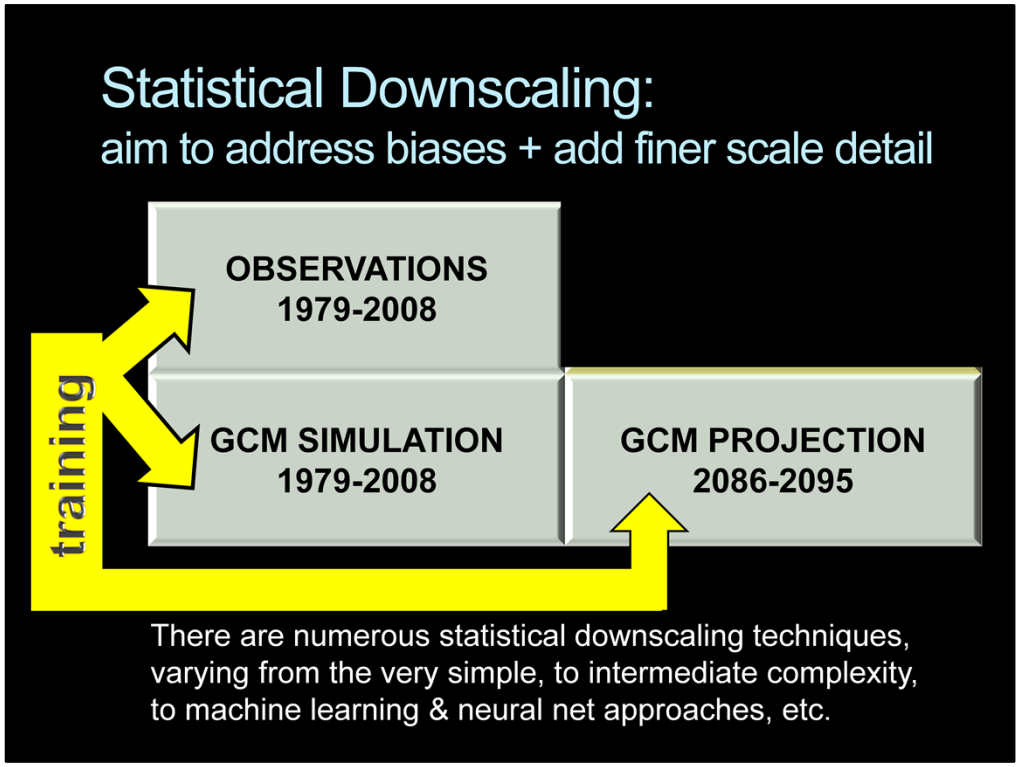
Yet, one way a GCM's results can fall short of the expectations of end users is that GCM output is often on a too coarse grid to provide the desired level of local detail. For example, a GCM with 200km grid spacing can have a single grid cell that overlies ocean waters, beaches, mountains, and everything 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). Empirical Statistical Downscaling (ESD) is often employed in an attempt to account for model biases and to provide additional spatial detail. A statistical downscaling step, generally considered a "value-added" process, does however, introduce its own uncertainties and different

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.

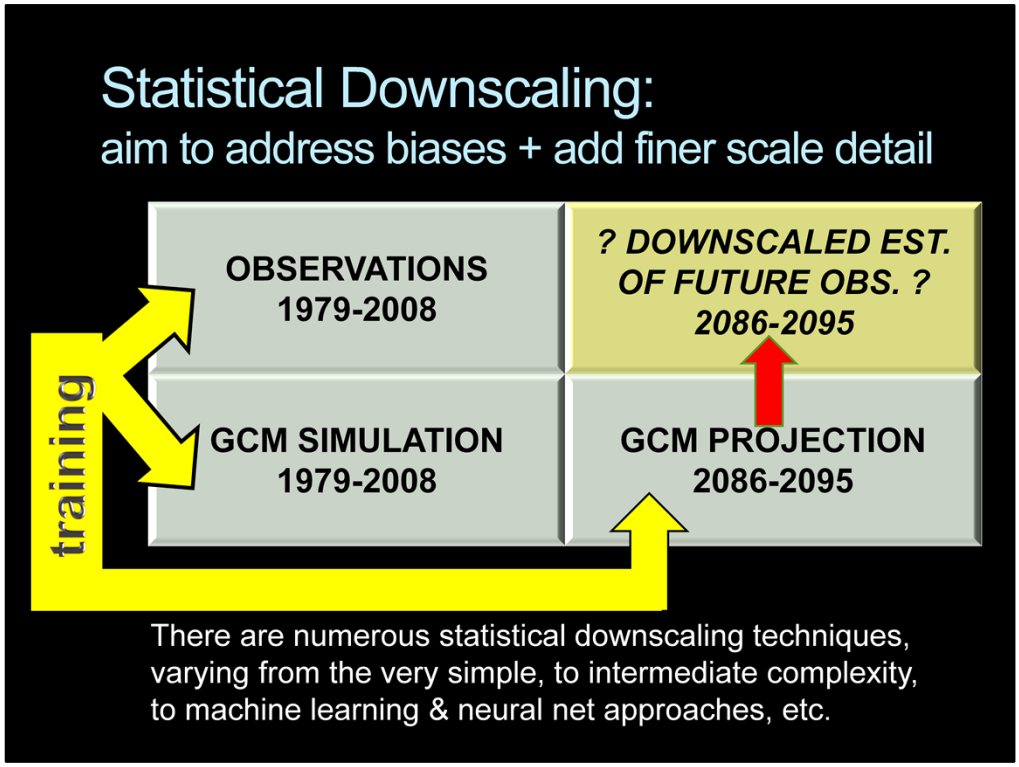


In this talk, we present an experimental design developed to isolate and quantify one particular source (but not the only source) of uncertainty arising from statistical downscaling of climate projections – the stationarity assumption.
The stationarity assumption is inherent to statistical downscaling of multi-decadal climate change projections, with the assumption being that statistical relationships 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 transfer 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’ framework, as described on the following slides.
The goals of this work include generating qunatitative infomration about ESD method performance that can…

(a) aid users of ESD data products, so that they might make better informed decisions when determining which ESD projections may or may not be appropriate for their climate impacts application of interest.
(b) aid ESD method developers to assess some of the strengths and weaknesses of their methods – information that can help guide them as they develop improved versions of their ESD methods.



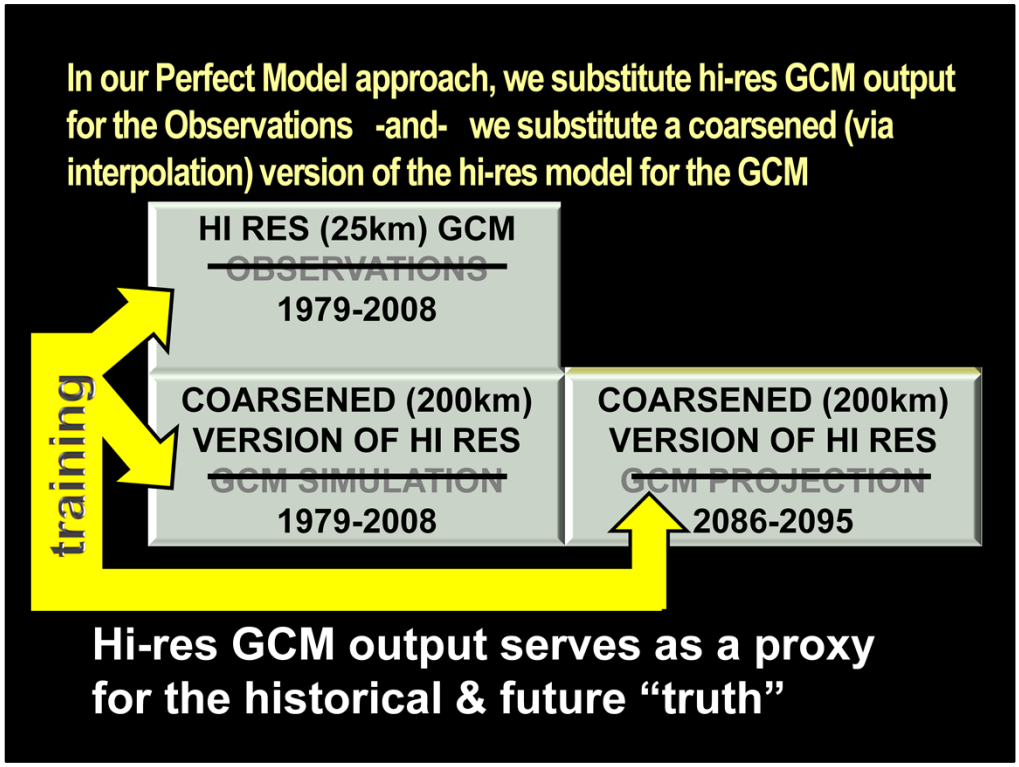
A summary of how ESD methods work…
Here we illustrate schematically the way in which our 'perfect model' framework isolates and quantifies the extent to which the stationarity assumption holds. We will do so by first describing one typical way that statistical downscaling is implemented in practice. We will contrast that with our perfect model framework, in subsequent slides.
All typical applications of ESD start with three types of data sets - one data set containing observations at the locations of interest for a historical period (here 1979-2008) and a companion dataset of GCM model results for the same time period. One also has a GCM data set from a future time period. The aim is to address GCM biases and other shortcomings such the downscaled output will more closely resemble what observations for the period of interest would look like.
During the 'training step', statistical methods are used to compare some combination of these 3 datasets (choice varies among methods, some use 2, others use all 3), and in the process downscaling transfer functions are derived that express relationships between the training datasets. 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 transfer functions 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 (transfer 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 are used in many climate impacts studies used in decision support 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.

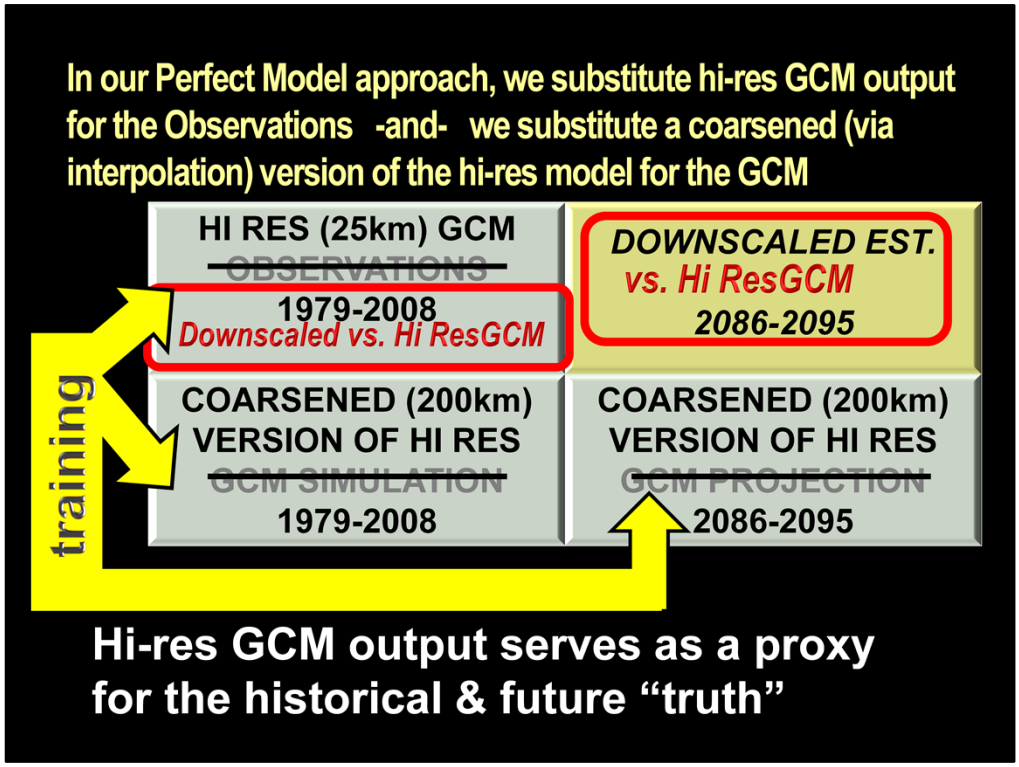
NOTE: Numerous statistical downscaling techniques exist. They span a wide range of complexity and can differ somewhat in the combination of data sets used in the training step. The schematic outline provided here represents most ESD techniques. What all ESD methods have in common when used to produce 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 combinations of 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 slide.

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.

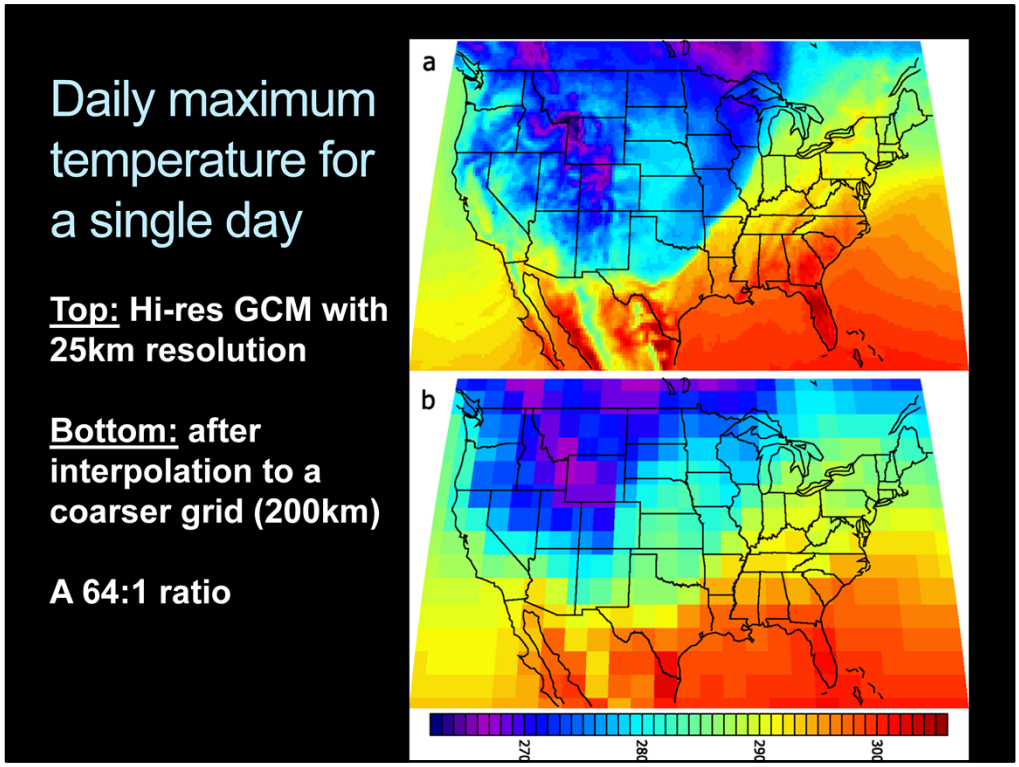
In our framework, the statistical downscaling training step that uses as input the 1979-2008 high resolution GCM data set as predictands and a coarsened version of itself as predictors yields transform functions that, in effect, attempt to recover the finer scale details and information lost as a byproduct 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.



This slide shows a snapshot of a single day's daily maximum temperatures over the geographic domain used in this study. The upper panel shows the High resolution GCMs field on a day when a strong cold front extended from the Grate lakes to Texas.

The lower panel shows the results of interpolating the High Res field to a coarser 200 km grid (64 grid points from the upper panel fit in each of the lower panel's grid cells.)

While the area averages of the two fields are essentially identical, we can point two 3 ways in which sizable differences (biases) are found at smaller scales where strong horizontal gradients exist.

(1) Visual inspection reveals that there are sizable differences along the cold front, as the interpolation greatly smoothed values in that part of the lower panel. This is transient… those same points wouldn't be expected to exhibit as large biases the next day.

(2) Persistent biases exist in areas a steep topography, as the atmospheric lapse rate leads to colder mountain peaks and warmer valleys in the high resolution field than in the smoothed field. The nature of these biases remains relatively similar year round.

(3) Strong gradients along coastlines can lead to biases, though they tend to change sign with the seasons.

For illustration purposes, we examine the performance of three statistical downscaling methods (quantile mapping & relatives)

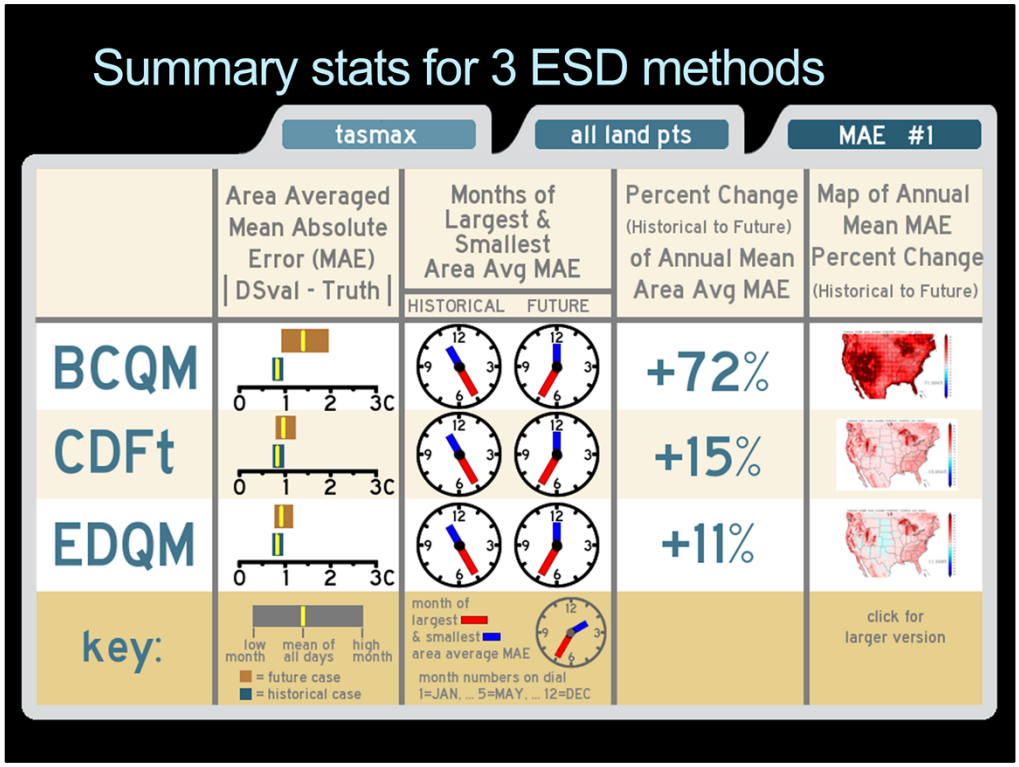
- **BCQM: Bias Correction Quantile Mapping** (train using historical obs & historical GCM; a stripped-down generic version)
- **CDFt: Cumulative Distribution Function transform** [Michelangli et al., 2009] (train with hist obs, hist GCM, & future GCM)
- **EDQM: EquiDistant Quantile Mapping** [similar to Li et al., 2010] (train with hist obs, hist GCM, & future GCM)

At GFDL we have tested several ESD methods, but for this short talk I will focus on just three that all can be classified as quantile mapping methods or close cousins (e.g., distributional methods). These were chosen more because they help illustrate our approach, rather than because they have a large 'market share' of ESD products used in climate adaptation studies.

BCQM = a generic, stripped down method that is one core part of several techniques used to generate multiple downscaled datatsets you may have downloaded, but typically developers have added features to the method that aim to enhance its performance beyond that of the simpler implementation we use here.

CDFT = a statistical downscaling code that is publically available (from the R Cran repository, Vrac et al.)

EDQM = our implementation of EDQM follows that of Li, Sheffield and Wood, but because it is not their code, we can't say that their code would produce identical results.



In the interest of time and space, I show here a very much condensed version of some of our Perfect Model results. Consider this an example of how we seek to present info in what some call a “Consumer Reports-like form”, in that it provides summary information users can scan and then decide it they wish to drill deeper.

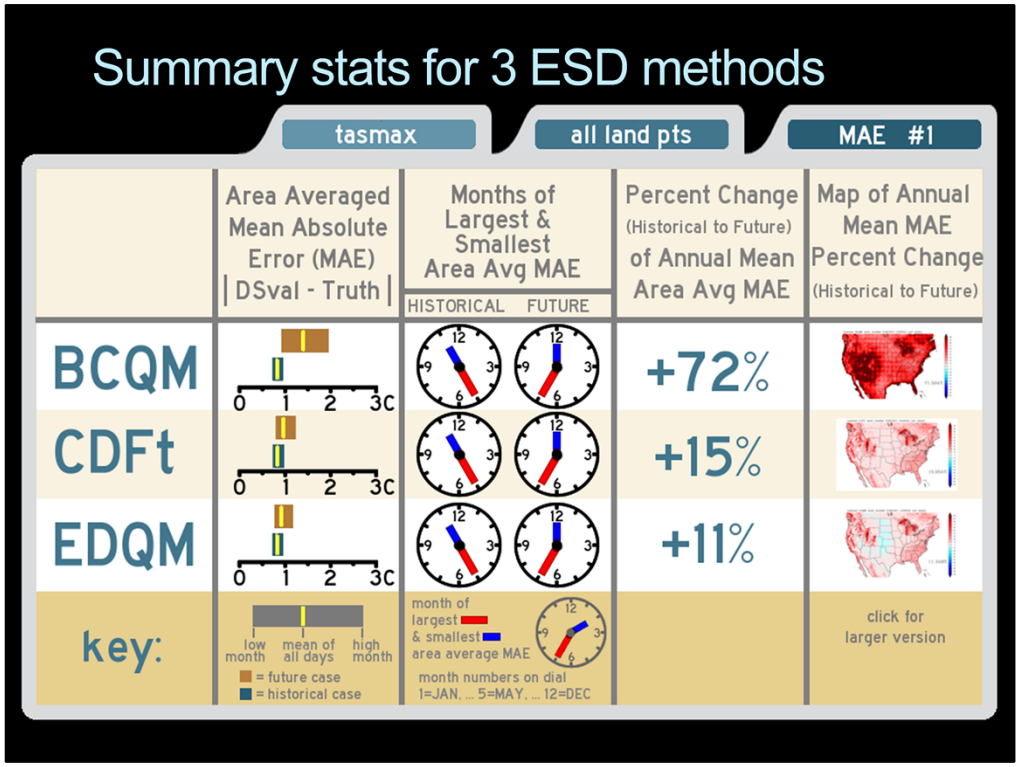
It’s also “Consumer Reports-like “ in tht we don’t aim to answer the question “Which ESD method is best for me to use?” Instead, we aim to provide information that others can use as part of their own determination of what method may be best suited for the demands of their application. Kind of like how for automobiles Consumer Reports lists statistics on gas mile, acceleration, braking distance, visibility, ‘fit and finish’, so that users can give weight to what matter most to them. (Unlike CR, we don’t provide overall summation scores or ‘best buy’ designations.)

The tabs at the top indicate that this slide reports on daily maxiximum temperature (tasmax) results. All the land grid points were considered when generating these results (no ocean points). And the results presented on this slide are all based on a Mean Absolute Error (MAE) metric computed as | (Downscaled Value) – (“Truth”) |

First column: statistical downscaling method names

Second column: Two horizontal bars per method. Top bar: MAE for the future time period, averaged over the area of interest (here, all land points). Bottom bar: MAE for the historical period. For each horizontal bar, the yellow vertical line represents the Area Avg Annual avg MAE… the left (lower) edge represents the lowest monthly MAE value of the 12 months, and the right edge represents the largest monthly MAE. So, the left to right extent of the bar gives a measure of the seasonal variations of the MAE statistic. If the historical and future MAE bars are indistinguishable, the stationarity assumption holds. However, for all 3 methods, the future MAE bar is shifted to the right, indicating a reduction in skill (larger downscaling errors), and hence we see that the stationarity assumption does not hold perfectly, especially for the BCQM method.

--continued--- >>



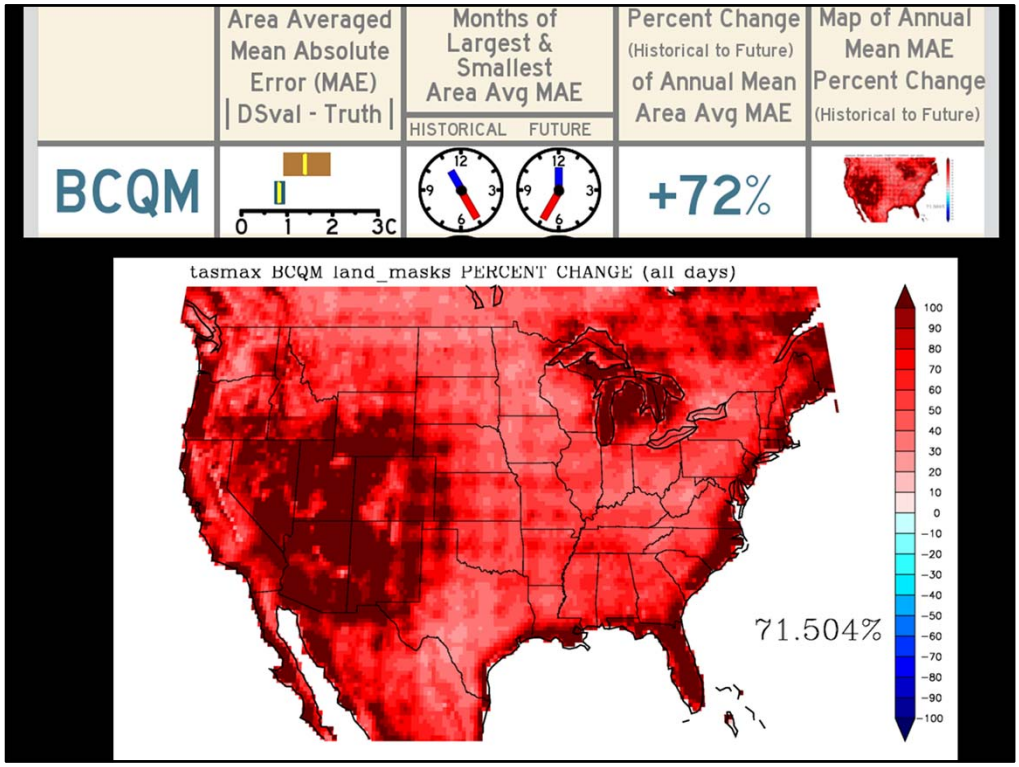
(slide repeated – explanation continued)

Third column: To provide more info about the seasonality of MAE variations averaged over the spatial domain of interest, we appropriate a clock face, but 1 thru 12 indicate the months Jan thru Dec, not hours. The short blue hand points to the month with the smallest MAE and the long red hand points to the month with the largest MAE. So, for all 3 methods, for the future time period the method performed best in winter (Dec) and had the largest downscaling errors in July.

Fourth column: Simple percent change of annual MAE computed over all grid points of interest (positive values indicate future MAE > historical MAE)

(i.e, How much do the two yellow bars (future and historical average avg annual avg MAE) differ in column 2 differ)

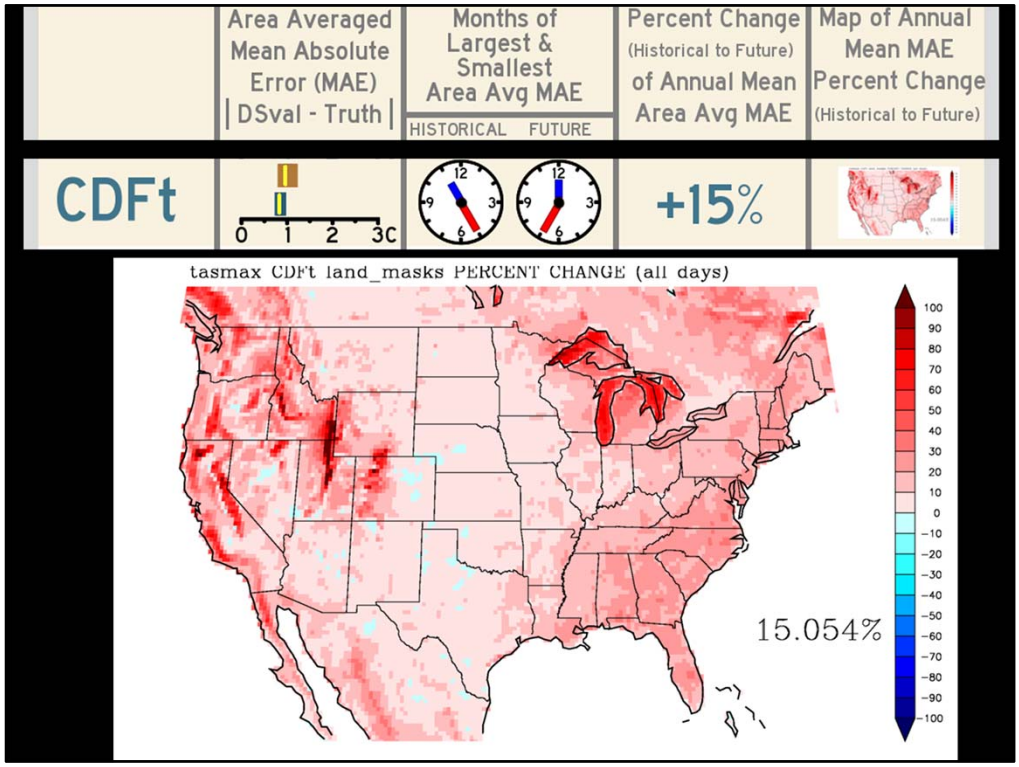
Fifth Column: Map showing spatial distribution of annual mean MAE statistic – depicts geographic variations in the method performance. (imagine clicking to see larger versions – next slides)



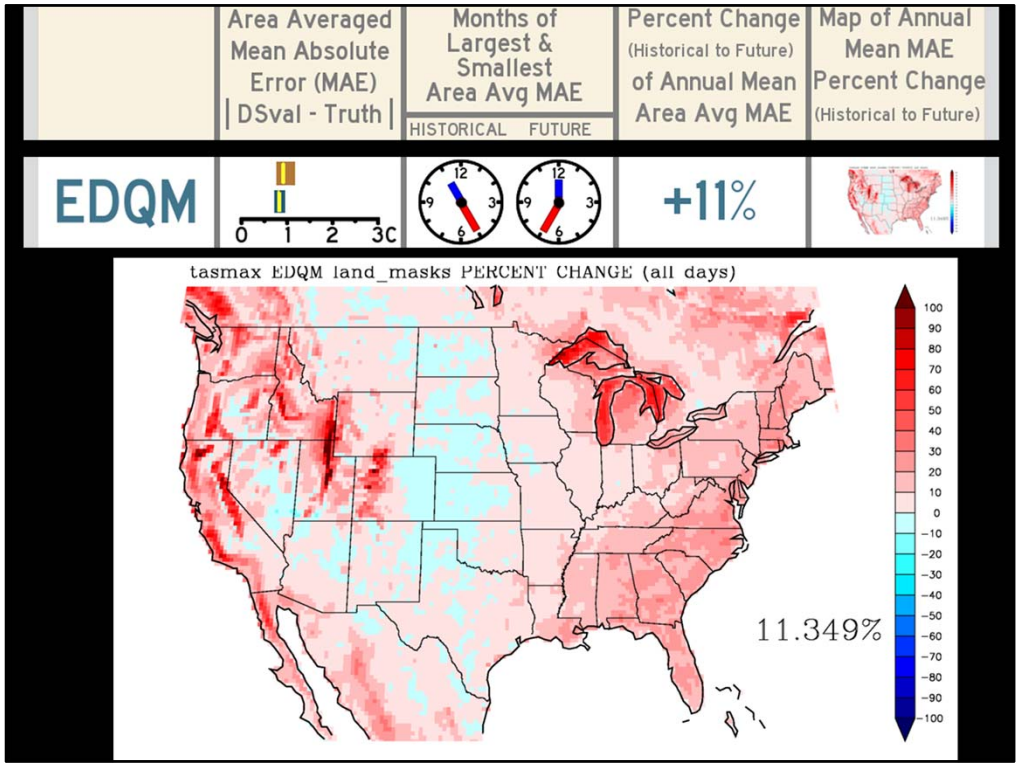
The BCQM method display the largest errors of the three tested here, indicating that the stationarity assumption does not hold well for the daily maximum temperature variable under conditions of sizable climate change. (In the RCP8.5 late 21st century case we use, the future averages about 7 degrees warmer than the historical, so it is not a subtle test in that regard).

Areas where the downscaling MAE is saturated dark red (>100%) had MAE values more than double relative to the historical period. (The area average of 72% that appears in column 4 also appear on the figure)

Areas showing the largest problems include near-coastal regions (including almost all of Florida), the Great Lakes, and areas of steep topography in the western US.

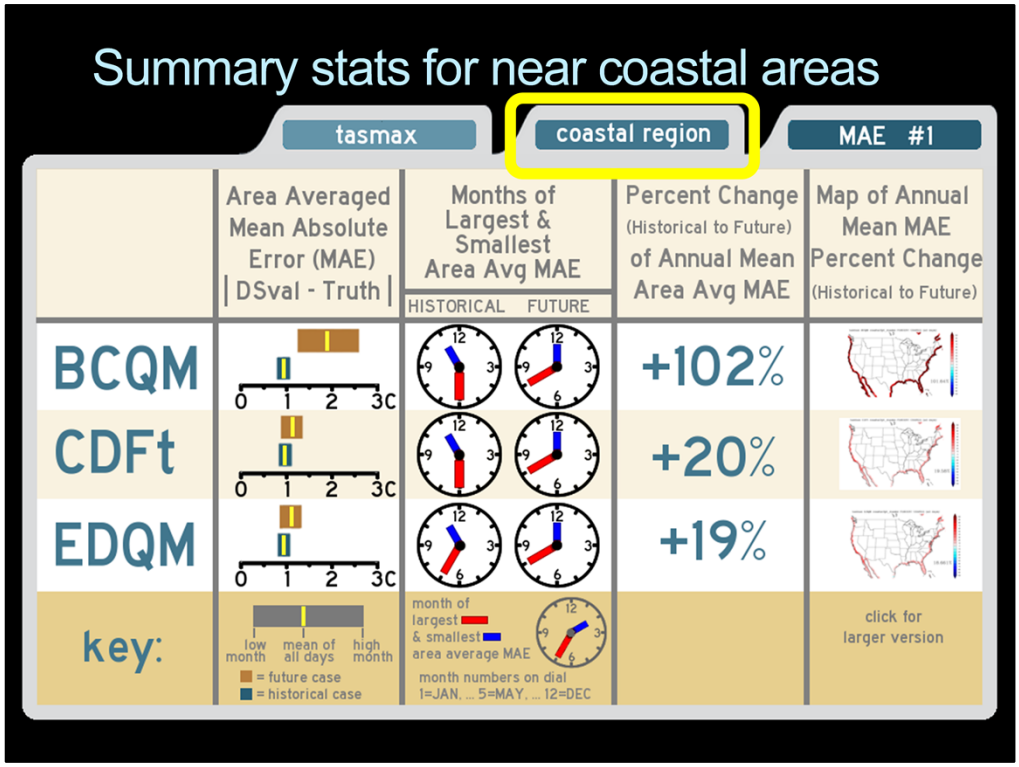


Though clearly performing better than the BCQM method overall, the CDFt method also shows marked geographic differences in method performance, with the central US showing little degradation in skill when applied to the future projection. The Great Lakes and some mountain areas seem problematic, but the coastal issue seen in BCQM appears to be less prominent.

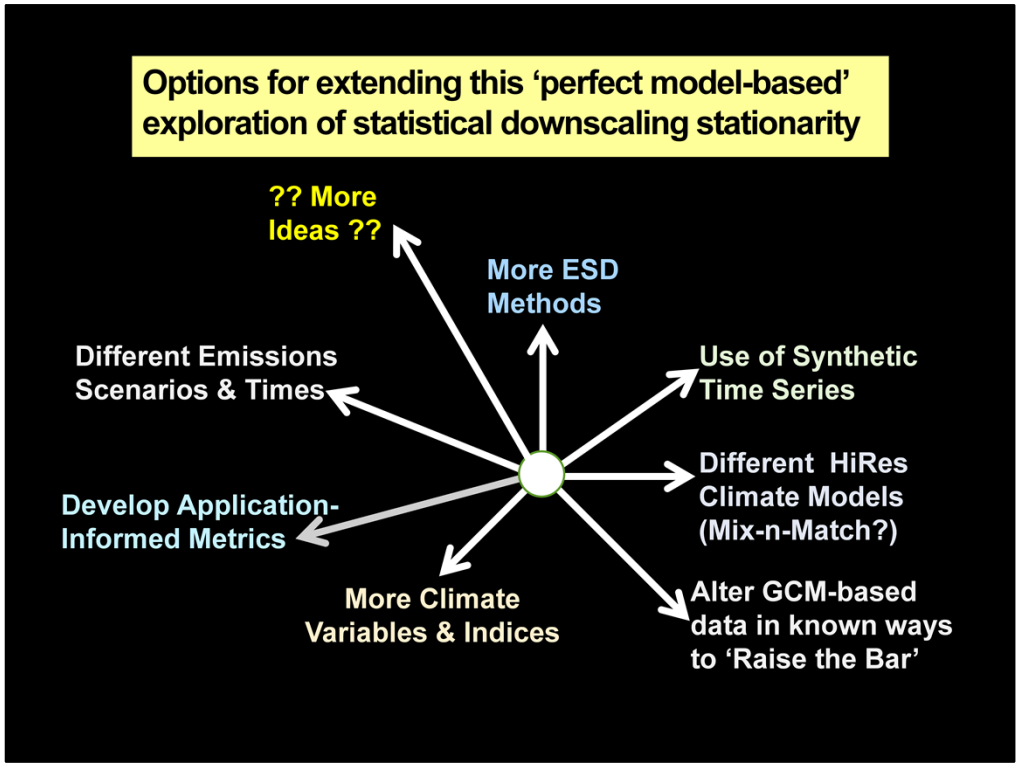


In general, the EDQM method yields results that are similar to (slightly better than) the CDTf downscaling method. Note that the transition from pale blue to pale red is the Zero line, so we see that for a swatch extending from Texas northward though the Dakotas, the stationarity assumption holds well when the EDQM method downscales daily maximum temperatures in this Perfect Model experimental framework.

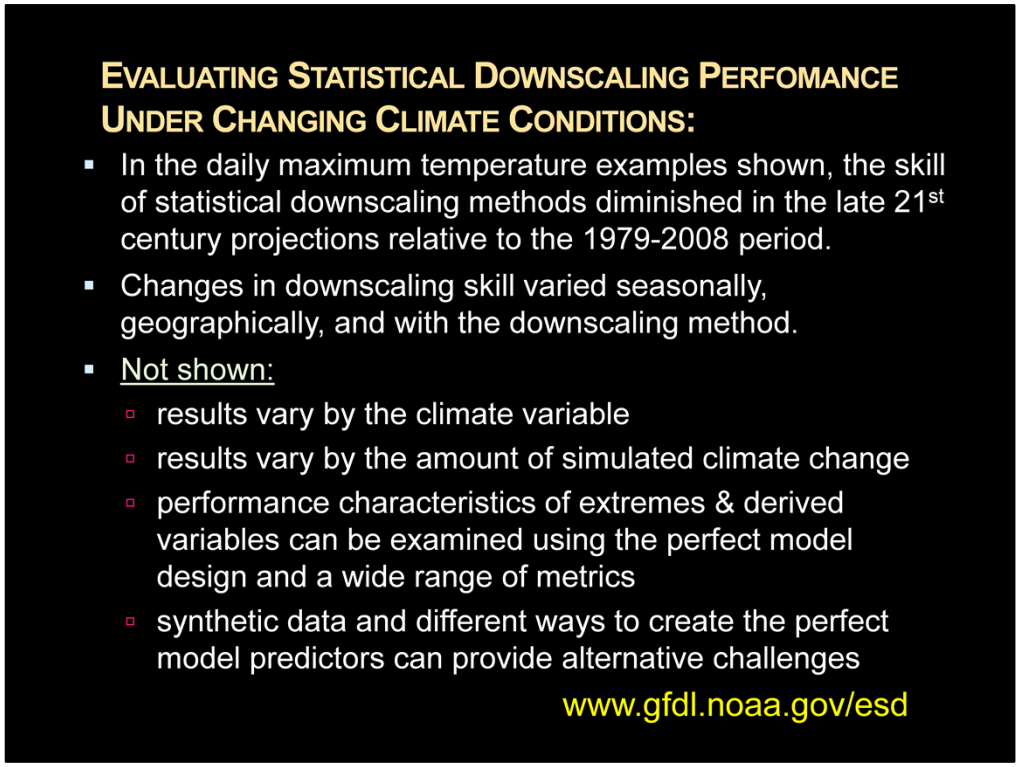
The CDFt and EDQM methods, while different mathematically, both use all three types of data sets (historical 'truth', historical coarse GCM, & future coarse GCM) in their 'training'. Training steps, whereas out version of BCQM uses only the 2 historical 'truth' and 'coarse GCM' data sets. Additional testing (not shown) suggests to us our variant of the BCQM method has difficulty producing reliable results when the future GCM data is outside the bounds of the historical GCM data used in the training step (think extrapolation vs. interpolation). This is consistent with the results that show all 3 of these methods perform similarly during the historical period (note similarity of the lower of each method's pair of MAE error bars found in column 2) but difference in future performance cane be quite marked.



This slide shows how the analyses one can do in this evaluation framework can focus on different geographic regions. We use “masks” (fields of values of either 1.0 or ‘missing value’ at grid points) to isolate different areas. In this example, we created a mask for near coastal regions – areas within 3 hi res ground pots (~75km) of saltwater (see small map insets in rightmost column). But one could just as easily create a mask for an individual state, river basin, etc.
We note that the column 4 percent change in the annual mean area avg MAE scores is largest in the coastal region (shown here) than when averaging over all land points (earlier slide).
-- -
I hope you can envision how this approach of generating and displaying information on ESD method performance could be applied to…
+ different climate variables (e.,g., daily precipitation, winds, or derived quantities like growing season length, number of consecutive dry days, etc.)
+ different geographic regions
+ different performance metrics (here we looked only at MAE, but a wide range of metrics can be used, such as exceedance statistics, measures of temporal and spatail variations, statistical comparisons of distributions, etc., etc.)
In effect, one can create many different Consumer Reports-like tables that would aim to touch on difference aspects of statistical downscaling performance and the validtiy of the stationarity assumption, with different tables being of more or less interest to an individual user based on his/her application of interest.



The Perfect Model Framework described here (and in a manuscript being submitted to the journal Cimatic Change, target: late-spring 2015) can be expanded and enhanced in a number of ways, several of which are alluded to in this slide.
The GFDL ESD Team welcomes your ideas and feedback about potential 'next steps'. At its inception, the Perfect Model approach to testing ESD method performance was adopted by the now defunct NCPP effort (National Climate Projections & Predictions Platform) as being a tool that could become part of a scientific community-based effort to generate knowledge and information relevant to the assessment of the 'credibility' of downscaled climate projection used in so many climate impacts studies and decision-support applications. We are still interested in pursuing that sort of collaborative approach, and as such welcome hearing from statistical downscalers and analysts/diagnosticians who might be interested collaborating.
In general, collaborations with esd developers might follow one of 2 paths–
"Data to Code": we share some of our perfect model data sets (or new ones) with esd method developers, so that they could run their method thru the perfect model framework and share the results with us –or-
"Code to Data": those who might be interested in sharing their esd codes with us sand working with us to incorporate their methods into our workflow, so that we might run texts and share the results with them.
(Though our experiences thus far are limited, we have received feedback from some ESD developers indicating that what they've learned from the perfect model results is helping to guide their efforts to improve their downscaling methods.)
And collaborations with diagnosticians/analysts can similarly proceed in different ways, with the common goal being to enhance and expand the set of metrics and analysis techniques that can be applied to look at perfect model results in a manner that focuses on performance aspects that are relevant to particular climate impacts studies (the Develop Application-Informed Metrics' arrow above).



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' 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 tweak the perfect model experimental design (e.g., using output from different GCMs, use synthetic data having known characteristics, to alter the nature of the challenge being presented to the ESD methods and to thereby test a wider range of performance characteristics.