CATS CENTRE FOR THE ANALYSIS OF TIME SERIES

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Evaluating model performance: A weather-like example

Emma Suckling and Erica Thompson

Centre for the Analysis of Time Series, London School of Economics

www2.lse.ac.uk/CATS

Workshop on Understanding Uncertainty in Environmental Modelling

8th January 2014

Evaluating weather-like models

CATS CENTRE FOR THE ANALYSIS OF TIME SERIES

Where we have enough past data to use ‘objective’ analysis frameworks

- Sufficient archive of forecast-outcome pairs for statistical evaluation
- Past data used to calibrate forecasts, weight models etc…
- Evaluations of past performance can give some insight into predictive capabilities

Data can be precious

- Information contamination
- Sensitivity to data used, calibration parameters, evaluation metric

Interpreting output and analysis is key (see Dave’s talk on Thur)

- Understand data, assumptions, sensitivities (see also Lindsay)
- Understand the limitations of models and analysis (see Erica)
- Different kinds of uncertainty (see Lenny)
- Implications for experimental design

Evaluating weather-like models

What question?

- Spatial/temporal scales, variables/indices of interest
- What is the application, who are the users?

What motivation?

- Understand? Inform? Motivate?

What approach?

- Perfect model scenario? Idealised study? Real-world? Policy-relevant?
- Deterministic? Probabilistic?
- Dynamical models? Empirical models?

Can my model answer the question?

- Understand what’s important
- Evaluate the models capabilities and limitations

Evaluation checklist: Data

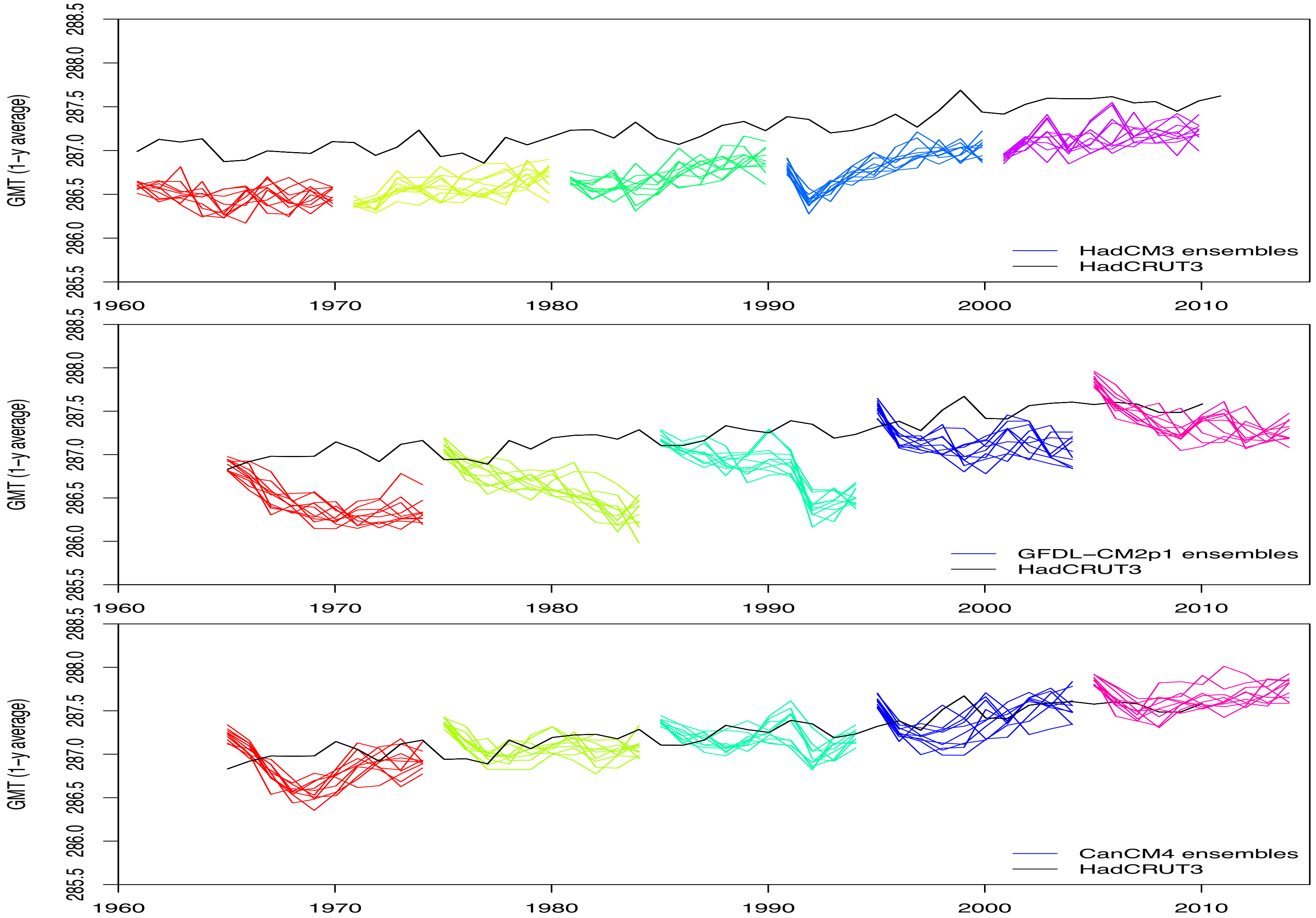
Data (models & observations)

- Availability, coverage, resolution, variables, different datasets, type of data
- What's interesting v. what's available
- Appropriate methods to robust evaluation will depend on all these

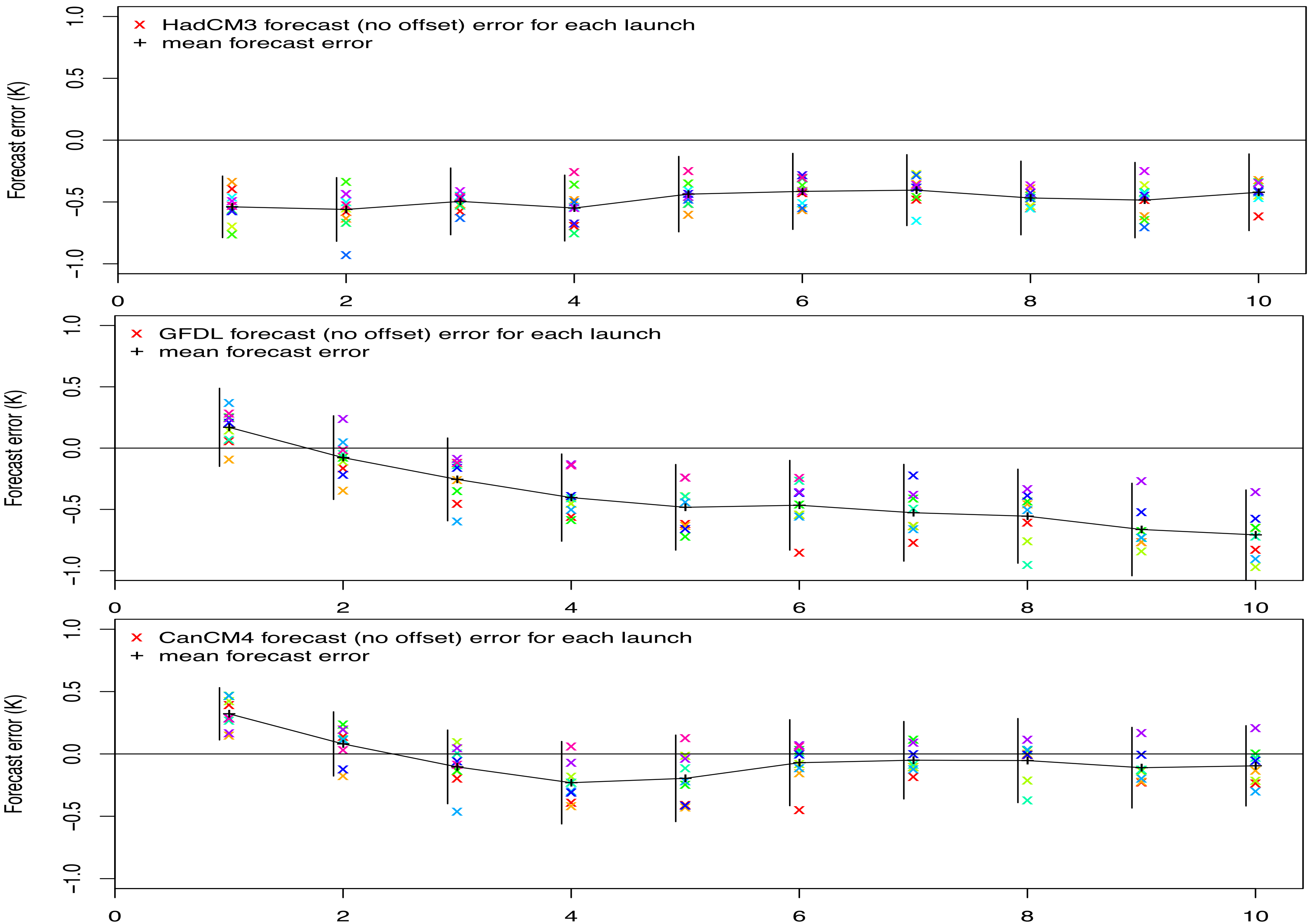
Look at the data!

- Before computing different indices, anomalies, metrics
- Get a feel for biases, uncertainties, ensemble spread etc…

Example: CMIP5 Decadal Hindcasts



Example: CMIP5 Decadal Hindcasts



Evaluation checklist: Data

Data (models & observations)

- Availability, coverage, resolution, variables, different datasets, type of data
- What's interesting v. what's available
- Appropriate methods to robust evaluation will depend on all these

Look at the data!

- Before computing different indices, anomalies, metrics
- Get a feel for biases, uncertainties, ensemble spread etc…

Notice when and how models fail

- Statistical evaluations show us when/how
- Physical insight can help us understand why

Evaluation checklist: Metrics

There are many performance metrics

Evaluation checklist: Metrics

There are many performance metrics

- Which to use?
- Avoid simply using ‘favourite’ or one that shows results in best light

Appropriate metrics depend on situation

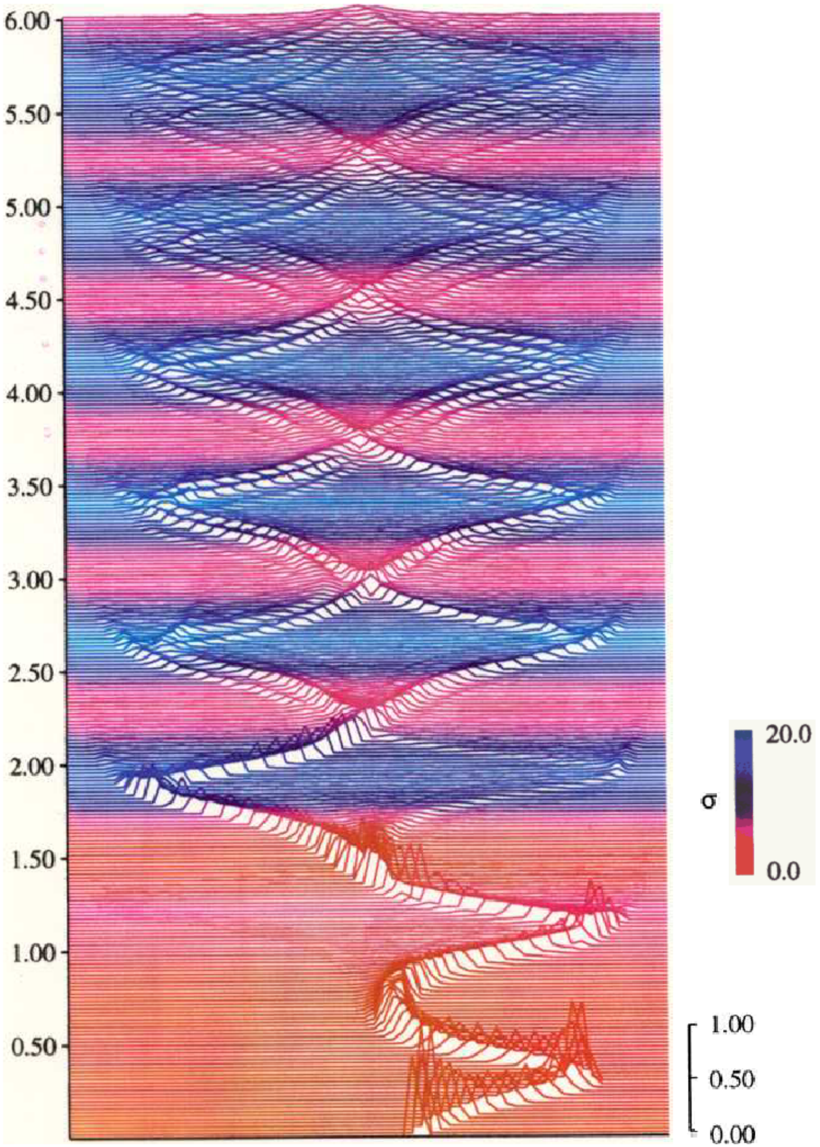
- Point forecasts, binary events, thresholds, ensembles, probabilities
- RMSE, correlations, reliability, ROC, Taylor diagrams, skill scores

Metrics can sometimes be misleading!

- RMSE of ensemble mean
- Spurious skill

Features of the data and model output interpretation determines which metrics are appropriate

Example: RMSE of ensemble mean



Lorenz 63: Probability of x with time

RMSE gives misleading estimate of skill

Features of the data and model output interpretation determines which metrics are appropriate

Example: Ensemble simulations

Ensemble mean

- Can be a useful diagnostic sometimes, but….
- Throws away information about uncertainties
- Do we expect the real-world to look like the mean?
- Can compute RMSE, correlation

Members represent only possible outcomes

- Zero probability mass between & outside individual members
- Some metrics rely on this interpretation but is it realistic/meaningful?

Members represent draws from a distribution

- Leads to a probabilistic interpretation (many ways to transform ensembles)
- Allows evaluation of output as would be used/interpreted by users
- Proper scoring rules are the only appropriate metrics

Proper scoring rules

<u>Proper:</u> All proper scores will give maximum reward to a forecast system when the forecasts and corresponding outcomes are drawn from the same distribution

Ignorance skill score: $S\left(p(y),Y\right)=\frac{1}{F}\sum_{i=1}^{F}-\log_2 P_i(Y_i)$

Continuous ranked probability score: $CRPS=\int_{-\infty}^{y}G\left(x\right)^2dx+\int_{y}^{\infty}\left[G\left(x\right)-1\right]^2dx$

Proper linear score: $PL=\int_{-\infty}^{\infty}P^2(x)\,dx-2P(Y)$

- **Even proper scores can rank two models differently!**
  - **When outcomes are drawn from a different distribution to model forecasts (imperfect model scenario)**

J. Bröcker and L. A. Smith, *Weather and Forecasting*, Vol. 22, 328 (2006).

Evaluation checklist: Benchmarks

Comparisons between empirical and dynamical models are useful

- Empirical models can serve as benchmarks for performance
- Could allow us to predict without knowing the laws of physics
- Allow quantitative comparison for regions/variables of interest
- Track improvements in dynamical models
- Can identify limitations of todays dynamical models
- Be used in combination or as cheaper alternatives

What makes an appropriate benchmark?

- Climatology, persistence, statistical models serve as common empirical benchmark models
- Is there a more appropriate model for the task?

Empirical v. Dynamical Models

Dynamical models ultimately expected to outperform empirical models

- Only dynamical models can capture the dynamics of the Earth System
- Do today's 'best available' models do so?

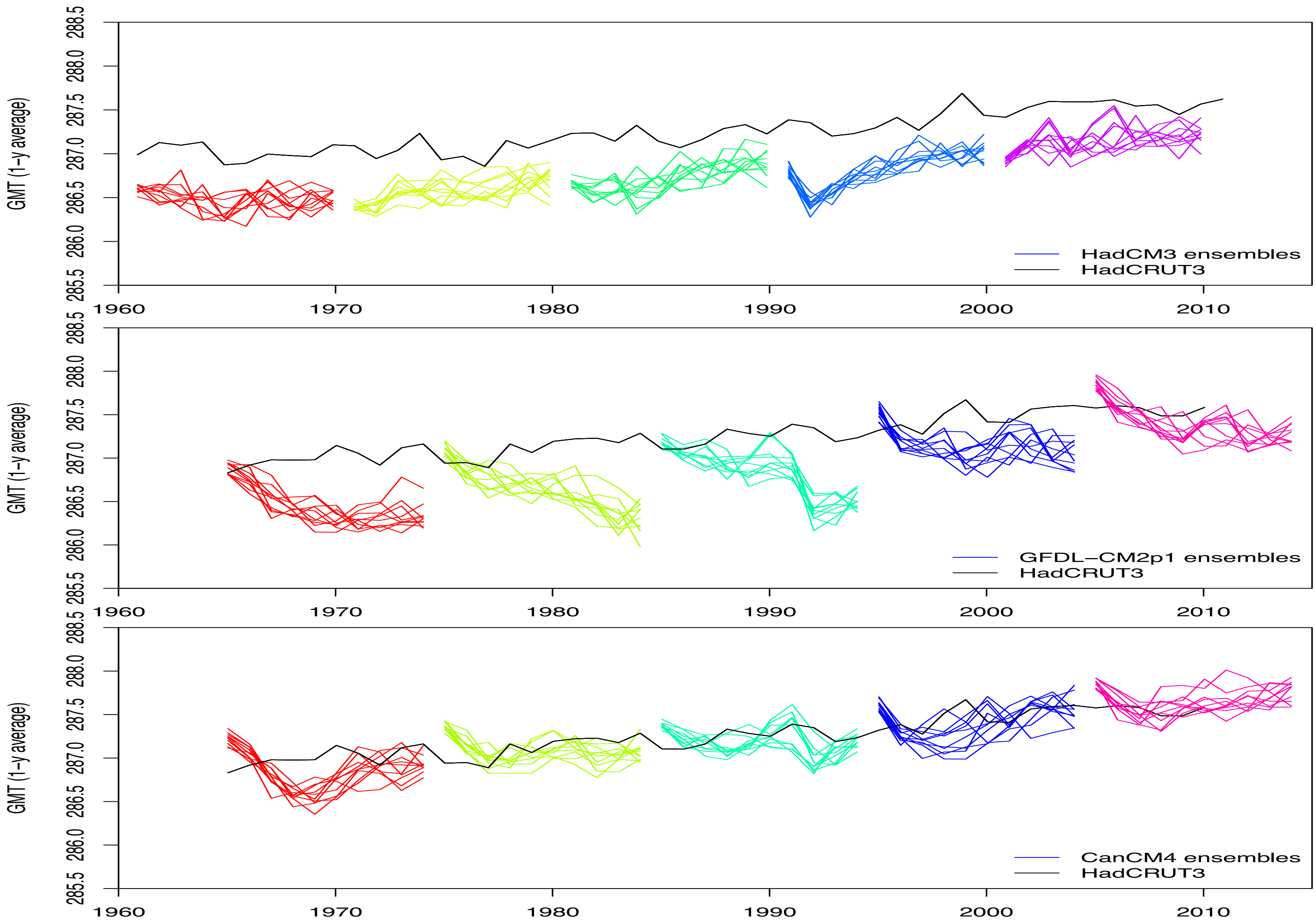
Example: Comparison of decadal hindcasts from CMIP5 and empirical models

- Focus on global and regional surface temperatures
- Annual lead times of 1-10 years ahead
- Methodology can be adapted to other variables or spatial scales of interest

Appropriate empirical benchmarks

- Dynamic Climatology used for comparison with initialized decadal hindcasts
- Empirical model can be adapted and refined for problem of interest

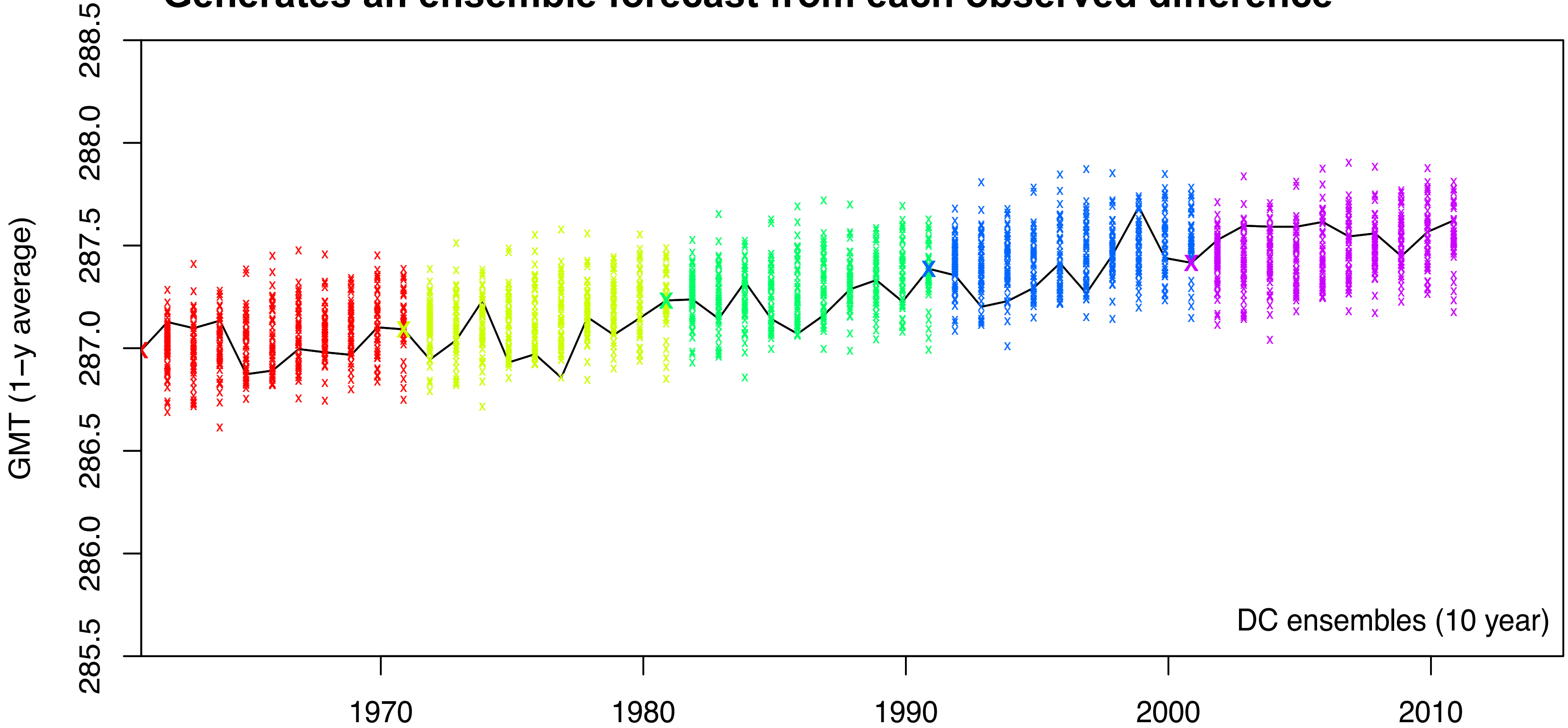
Example: CMIP5 Decadal Hindcasts



Dynamic Climatology Model

DC model uses differences in observed record

- Initialized to observations at each launch
- Generates an ensemble forecast from each observed difference



E. B. Suckling and L. A. Smith, *Journal of Climate, 26,23* (2013)
L. A. Smith, *Nonlinearity in Geophysics and Astrophysics*, CXXXIII:177-246 (1997)

Evaluation checklist: Interpreting output

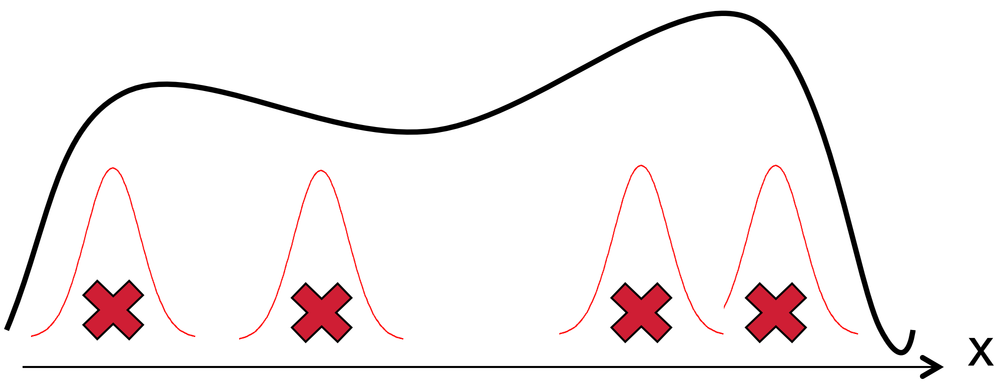
Kernel Dressing

The model-based component of the density, with *N* ensembles

$$p(y:x,\sigma) = \frac{1}{N\sigma}\sum_{i=1}^{N} K\left(\frac{y-(x^i+\mu)}{\sigma}\right)$$

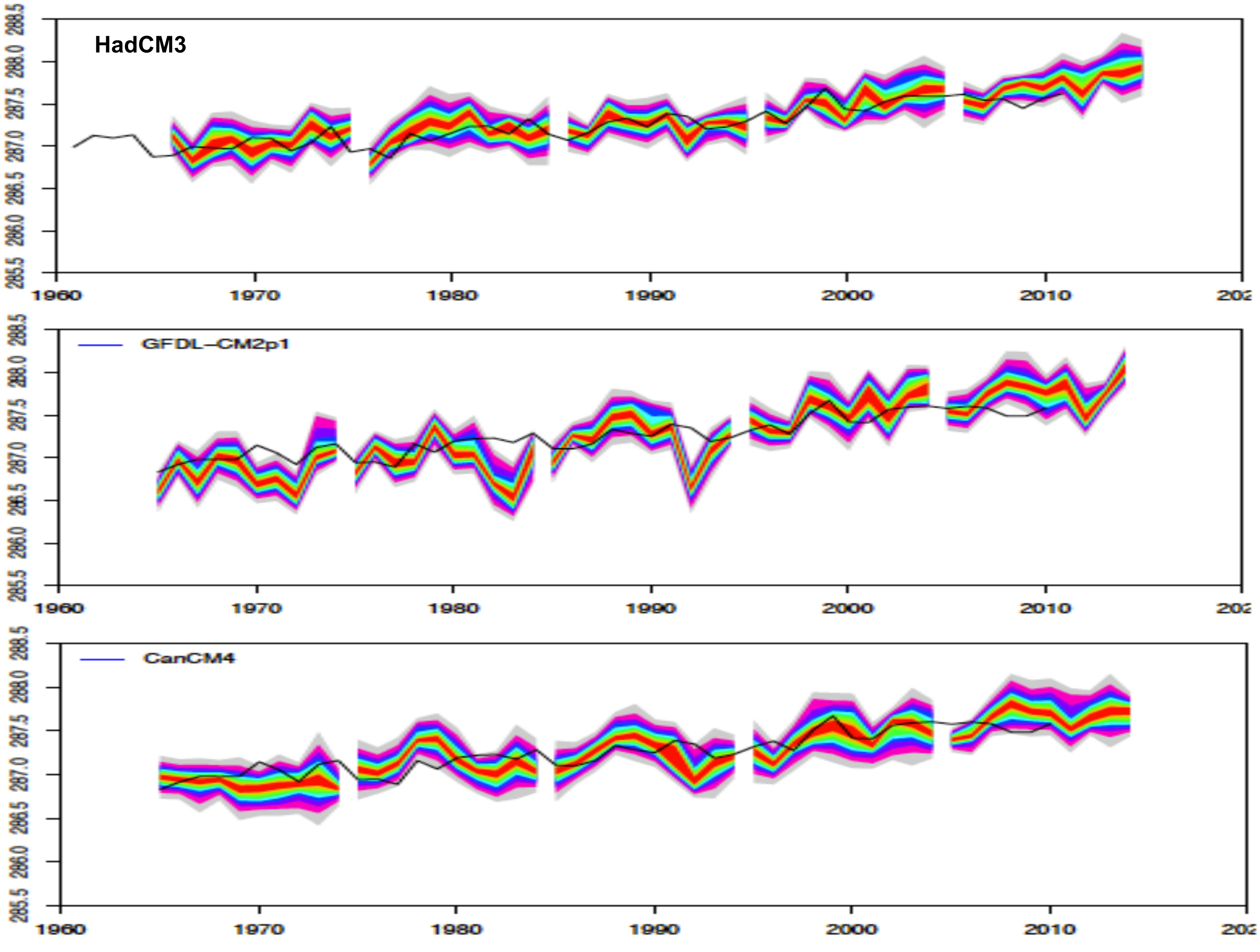
Kernel, *K*, is a normalized Gaussian

- **Parameters, μ (kernel mean plus bias correction) and σ (kernel width), are varied to minimise the skill score (Ignorance, CRPS, quadratic)**
- **Cross-validation method important in evaluation**

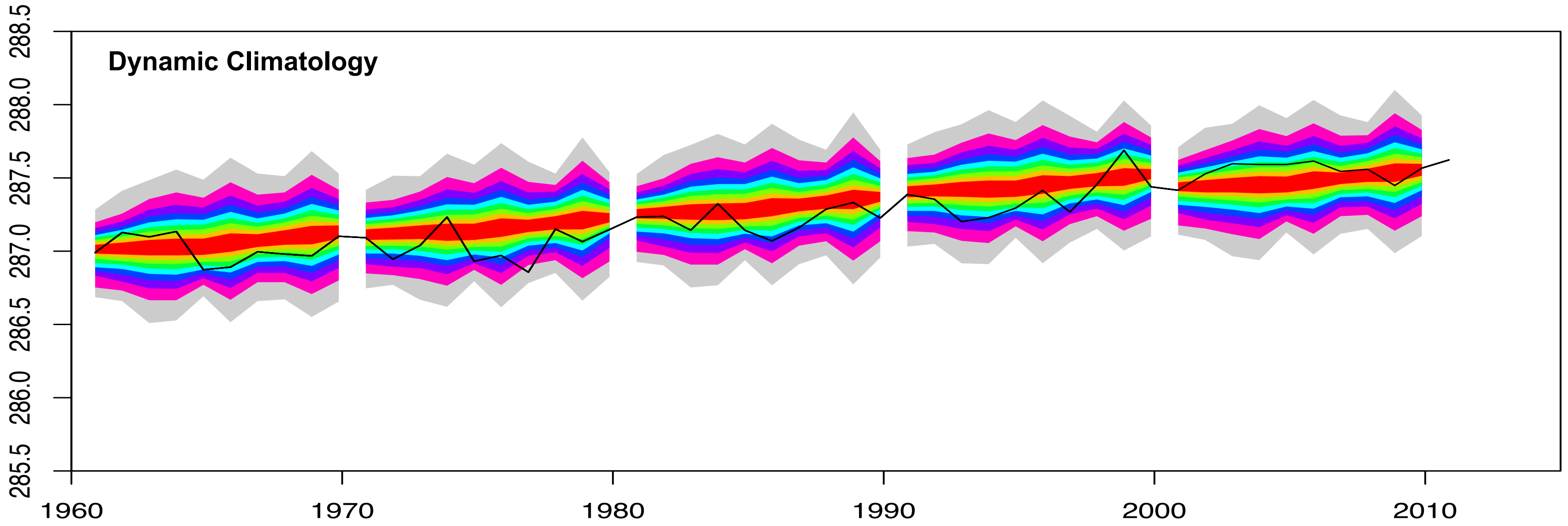


J. Bröcker and L. A. Smith, *Tellus A*, 60(4):663-678 (2007).

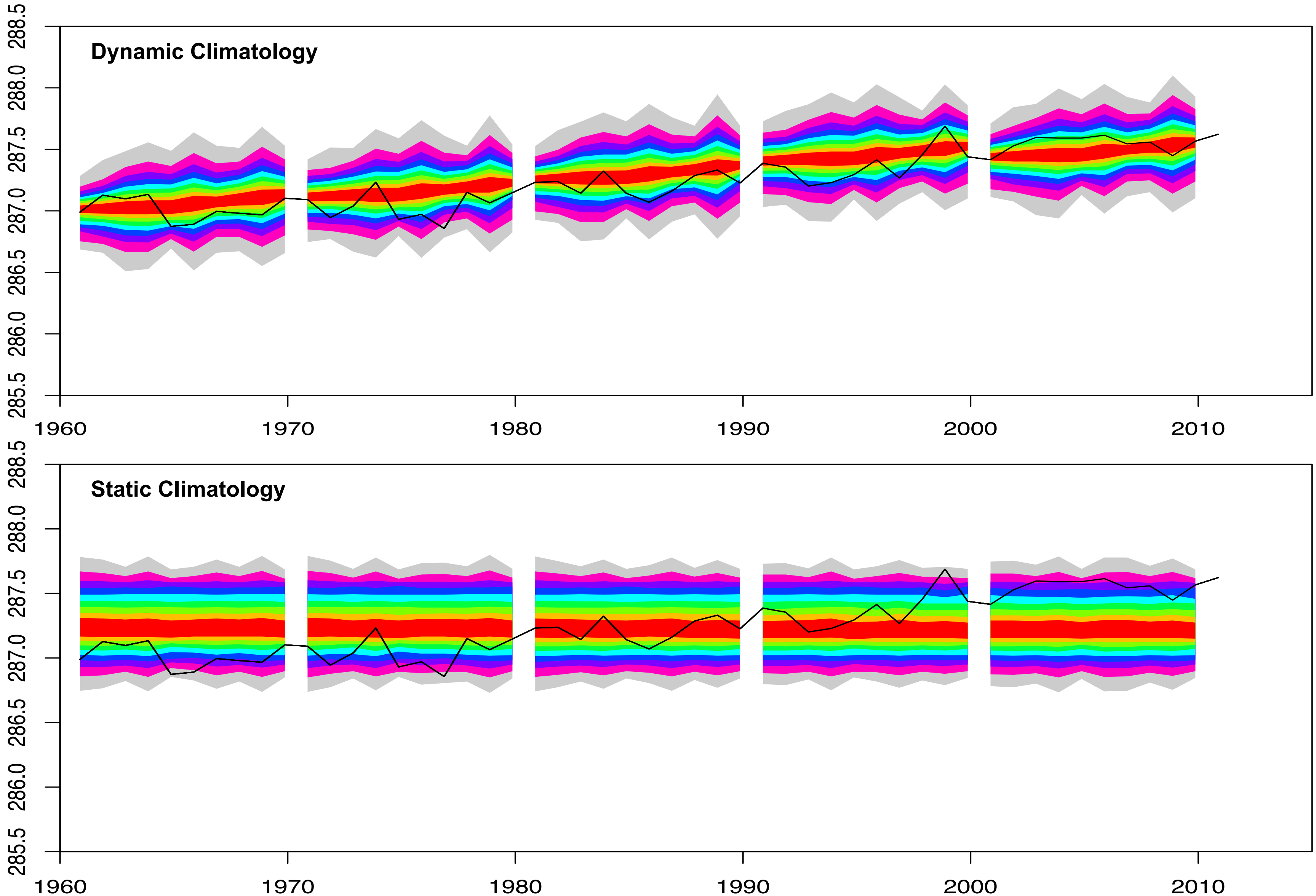
Example: CMIP5 Distributions



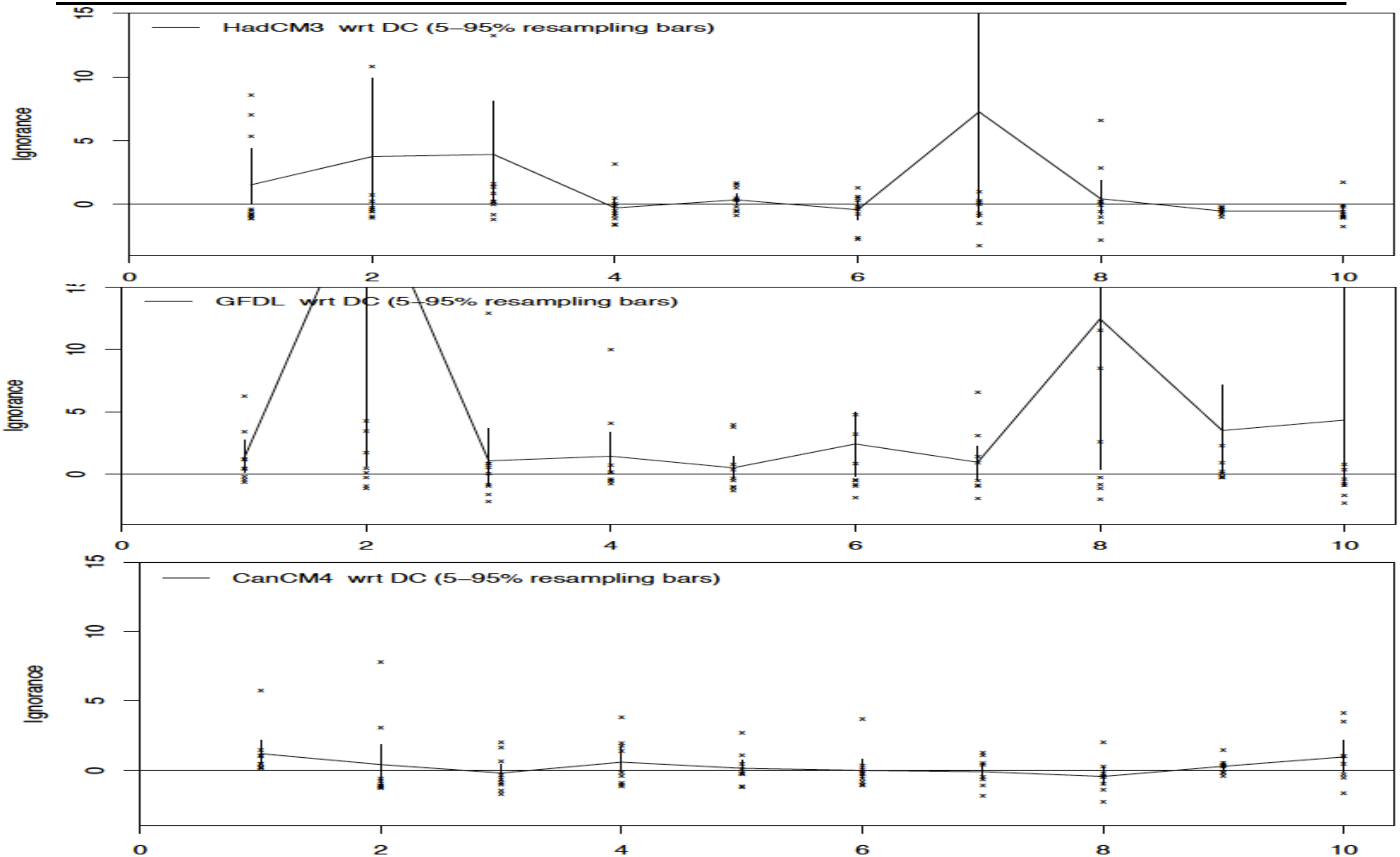
Dynamic Climatology Distributions



Dynamic Climatology Distributions



Example: Evaluation of Skill



E. Thompson, E. Suckling and L. A. Smith, An evaluation of CMIP5 decadal hindcasts, *in preparation*

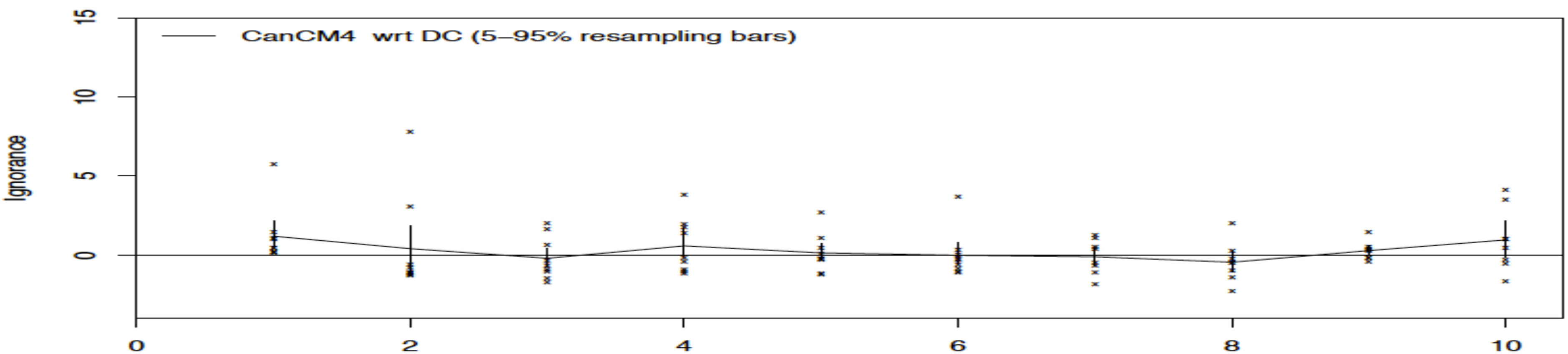
Example: Evaluation of Skill

DC empirical model and GCMs have comparable levels of skill

- What does this mean for prediction?
- Results robust to choice of metric, calibration and construction of DC model
- GCMs and DC show skill above static climatology
- Decadal hindcasts from CMIP5 models show improvement over previous projects (ENSEMBLES)

Can we combine dynamical and empirical models to improve utility?

- Blending forecasts could improve skill



E. Thompson, E. Suckling and L. A. Smith, An evaluation of CMIP5 decadal hindcasts, *in preparation*

Challenges

Experimental design & resource allocation

- Sample size – ensemble members, hindcast launches and period
- Consistent framework – model launch months, initialization

Calibration

- Dealing with model biases
- Information contamination (cross-validation)
- Interpreting ensemble output
- Combining/weighting models

Evaluation

- Appropriate empirical benchmarks
- Choice of metrics and indices

Sensitivity to different choices

Questions to consider

What assumptions have been made?

- Are they valid? Can I test? What could the consequences be?

What are the uncertainties?

- Which uncertainties are considered?
- What is the relevant dominant uncertainty?
- Aim to characerize, quantify, reduce uncertainty?

What is the sensitivity to different appropriate choices?

- Are results expected to change as new data comes in?

Does an in-sample evaluation give confidence out-of-sample?

Summary

'Objective' forecast evaluations are possible in weather-like case

- **Different choices in metric, calibration and ensemble interpretation can have an impact when data is precious**
- **Choose appropriate performance metrics for the task**
- **Empirical models provide useful benchmarks**

Consider assumptions, uncertainties and sensitivities

Combine information from statistical evaluations and physical insight

- **Combining information from empirical and dynamical models could be of value**
- **Statistical evaluations quantify capabilities and limitations of models**
- **Investigate when/where and why models fail**

Thank You!

e.suckling@reading.ac.uk