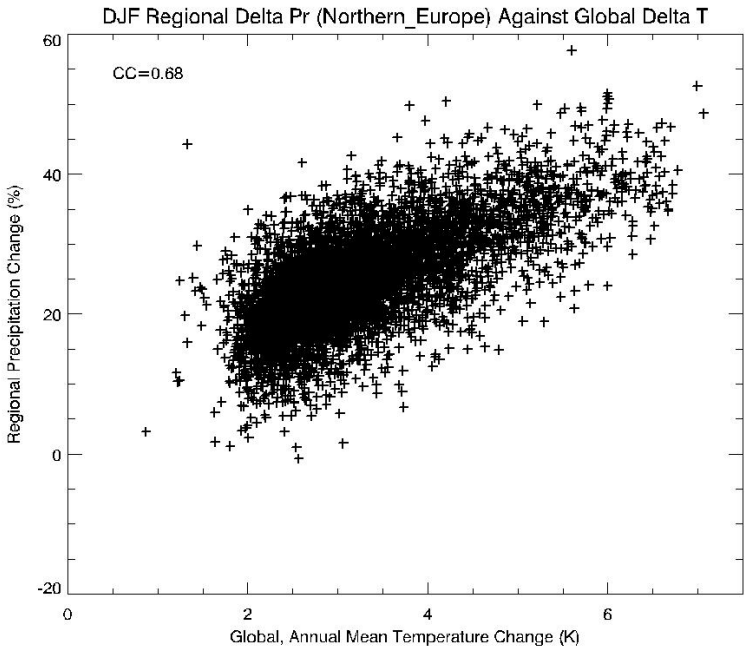
Challenges in the Interpretation of Ensembles: Why Good Statistical Methods Aren't Enough

Dave Stainforth

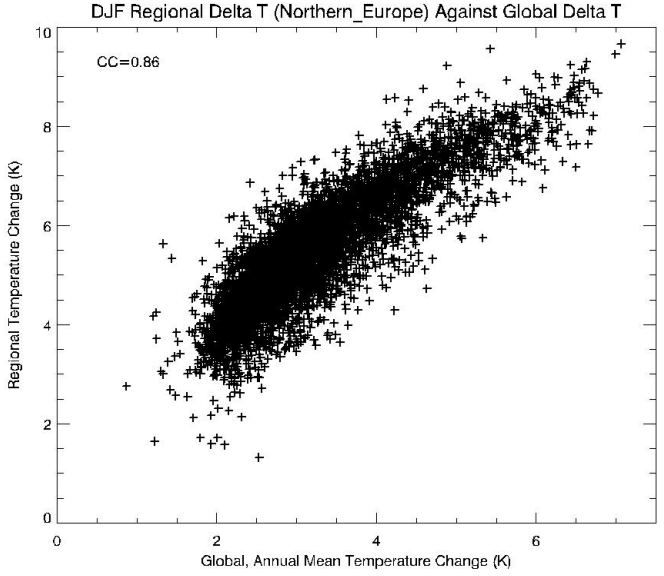
Centre for the Analysis of Timeseries and Grantham Research Institute on Climate Change and the Environment, **London School of Economics**.



Understanding Uncertainty in Environmental Modelling

LSE

9th January 2014



LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Issues

- Independence
- Model culling or weighting.
- In-sample ensemble analysis.
- Extrapolation.

Regional / Local Predictions
An Area of Significant Effort

UKCP09:

"The UK Climate Projections (UKCP09) provide climate information designed to help those needing to plan how they will adapt to a changing climate. The data is focussed on the UK,"

"UKCP09 provides **future climate projections** for land and marine regions."

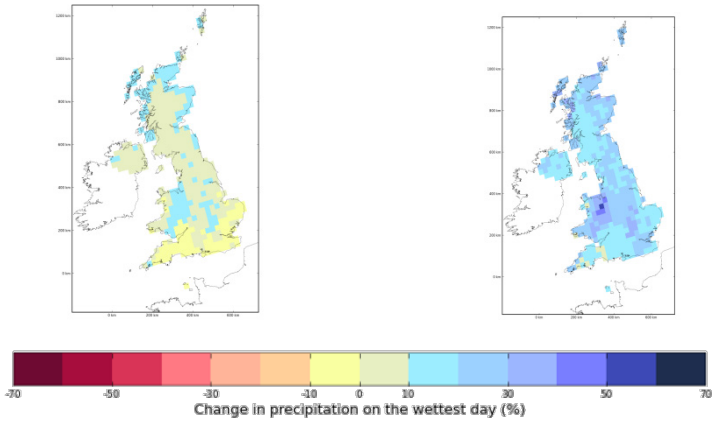
"They **assign probabilities** to different future climate outcomes. "

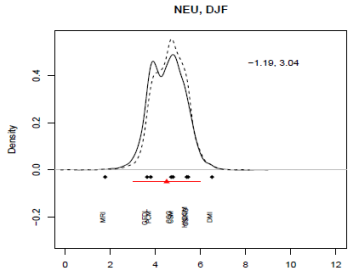
http://ukclimateprojections.defra.gov.uk

Change in Wettest Day in Summer Medium (A1B) scenario

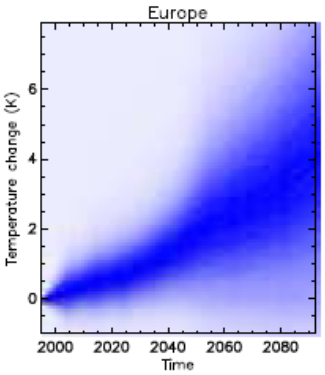
2080s : 67% probability level: unlikely to be greater than

2080s: 90% probability level: very unlikely to be greater than



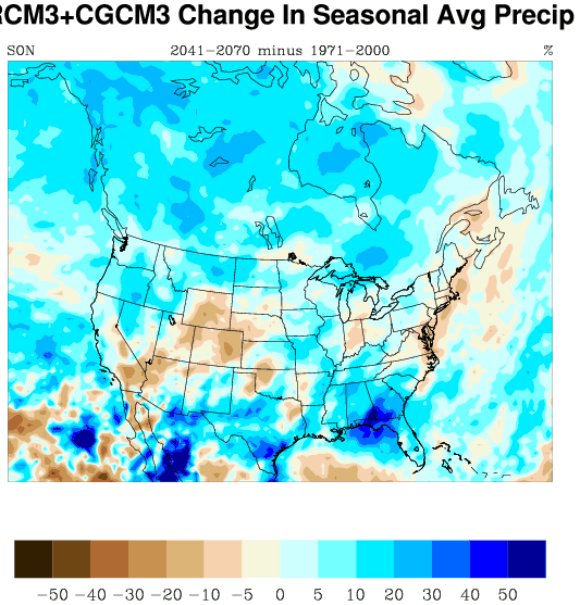


Tebaldi et al.., JoC, 2005



Stott et al.., GRL, 2006

NARCCAP:



The North American Regional Climate Change Assessment Program (NARCCAP) aims to "**investigate uncertainties in regional scale projections of future climate and generate climate change scenarios for use in impacts research**."

http://www.narccap.ucar.edu/about/index.html

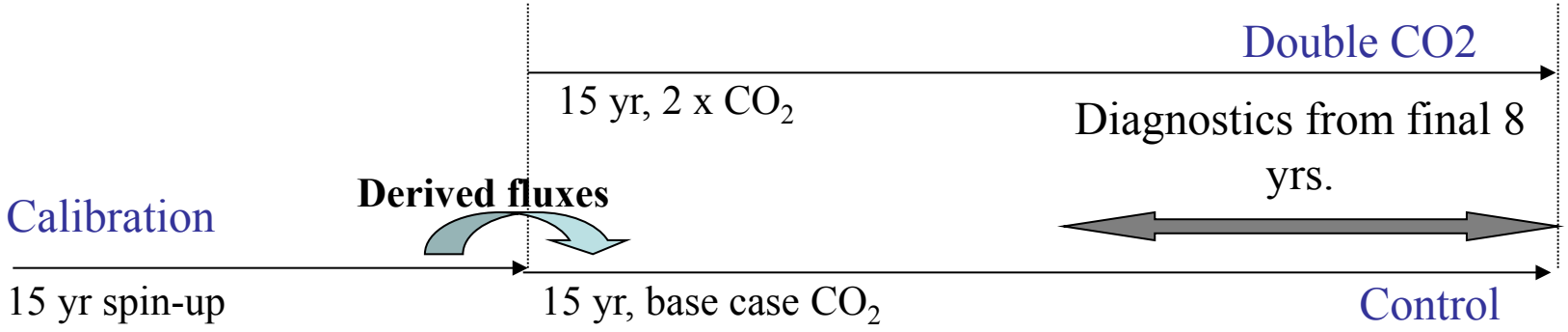
€7M European Call:

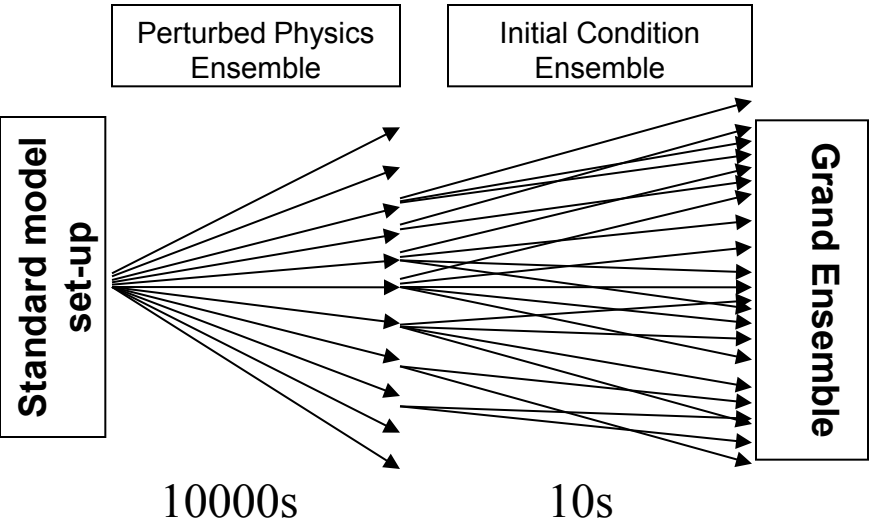
ENV.2011.1.1.6-1 "The proposed research activities should […] **quantify the impacts of climate change in selected areas of Europe** […] arising from a global averaged surface temperature change of 2°C from preindustrial level."

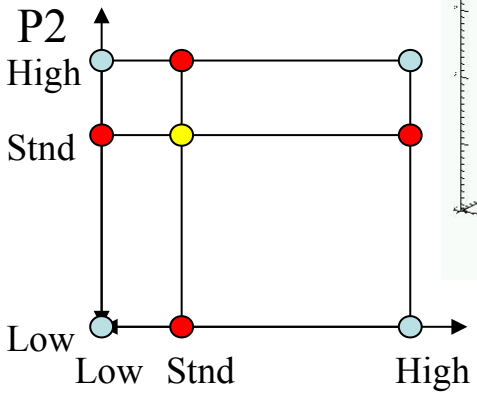
ftp://ftp.cordis.europa.eu/pub/fp7/docs/wp/cooperation/environment/f-wp-201101_en.pdf

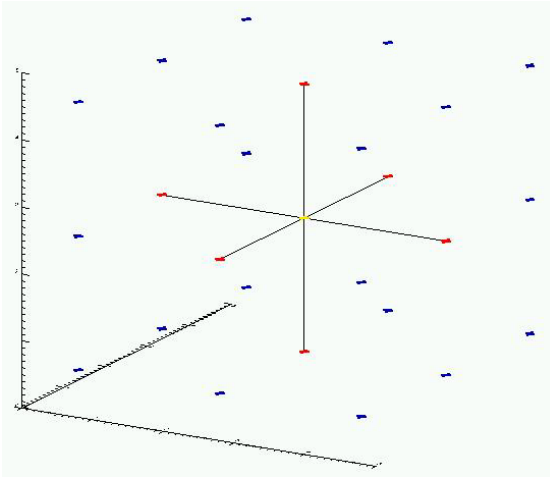
Climate*prediction*.net: The Slab Model Experiment

- Unified Model with thermodynamic ocean. (HadSM3)

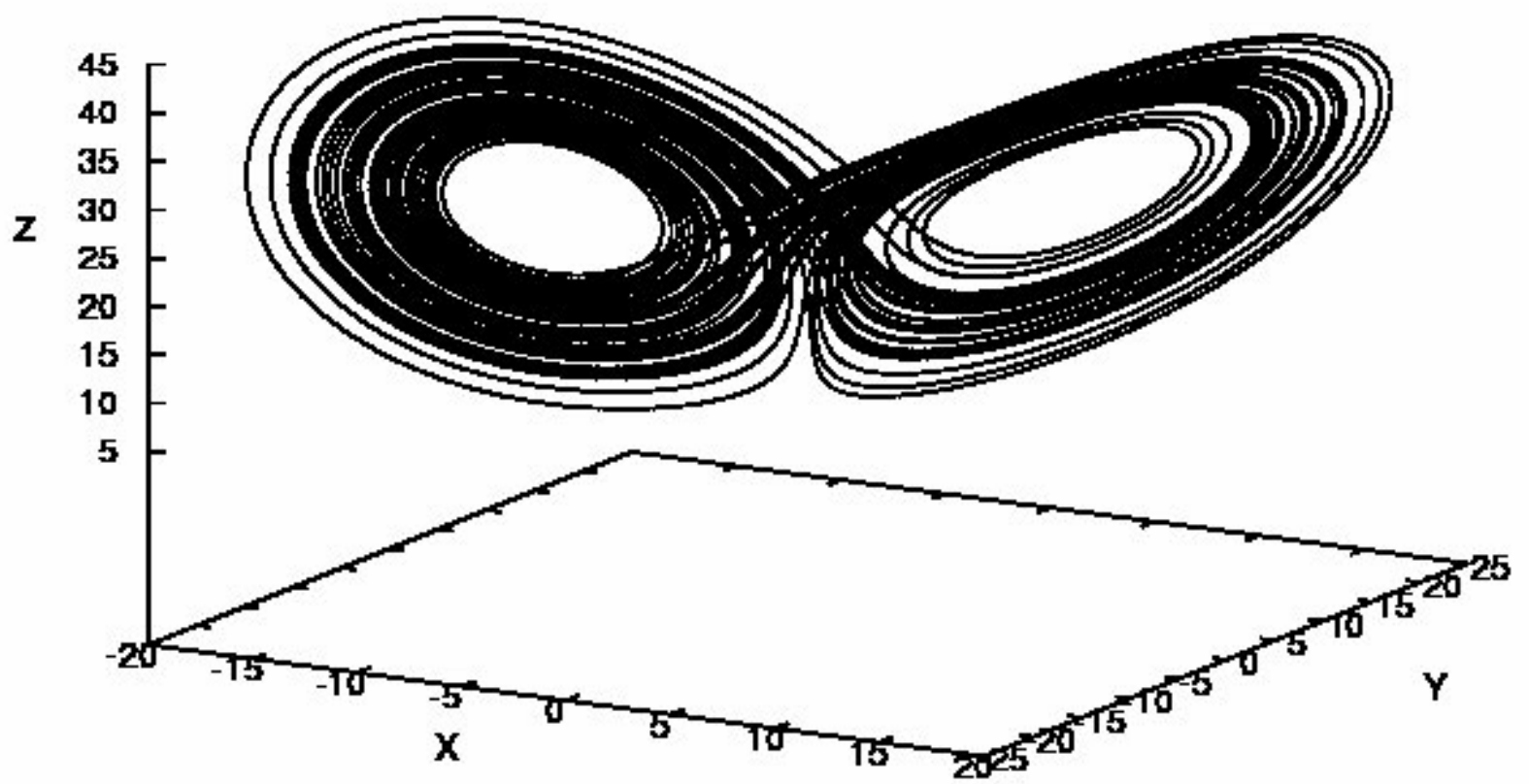




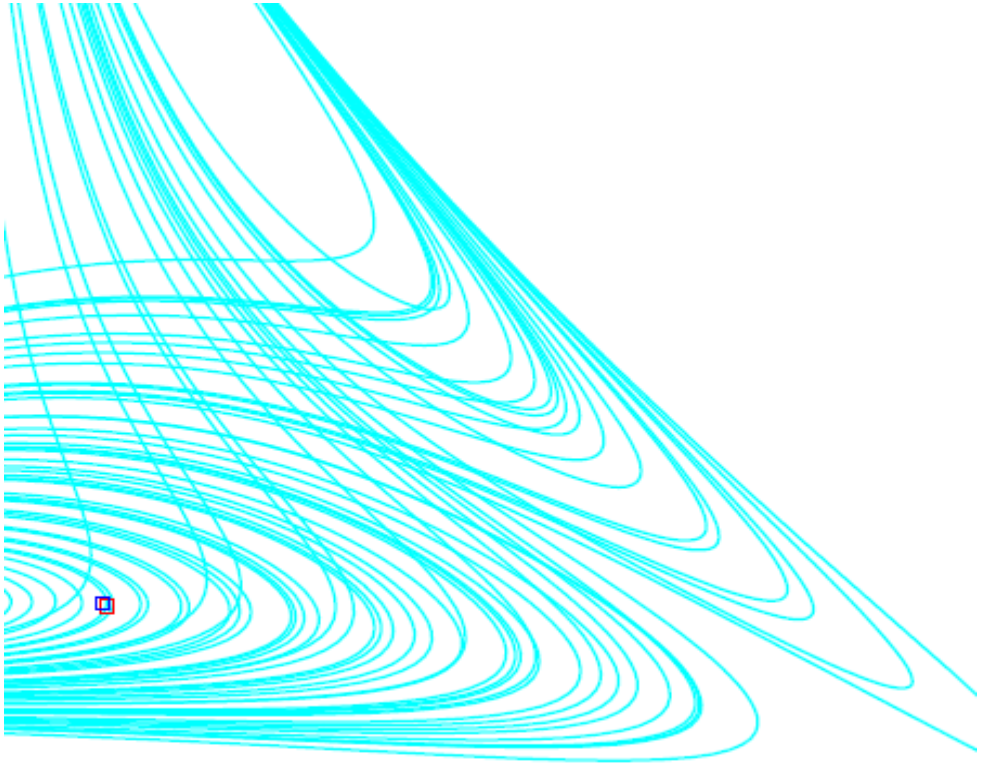




Lorenz 63 and the Butterfly Effect

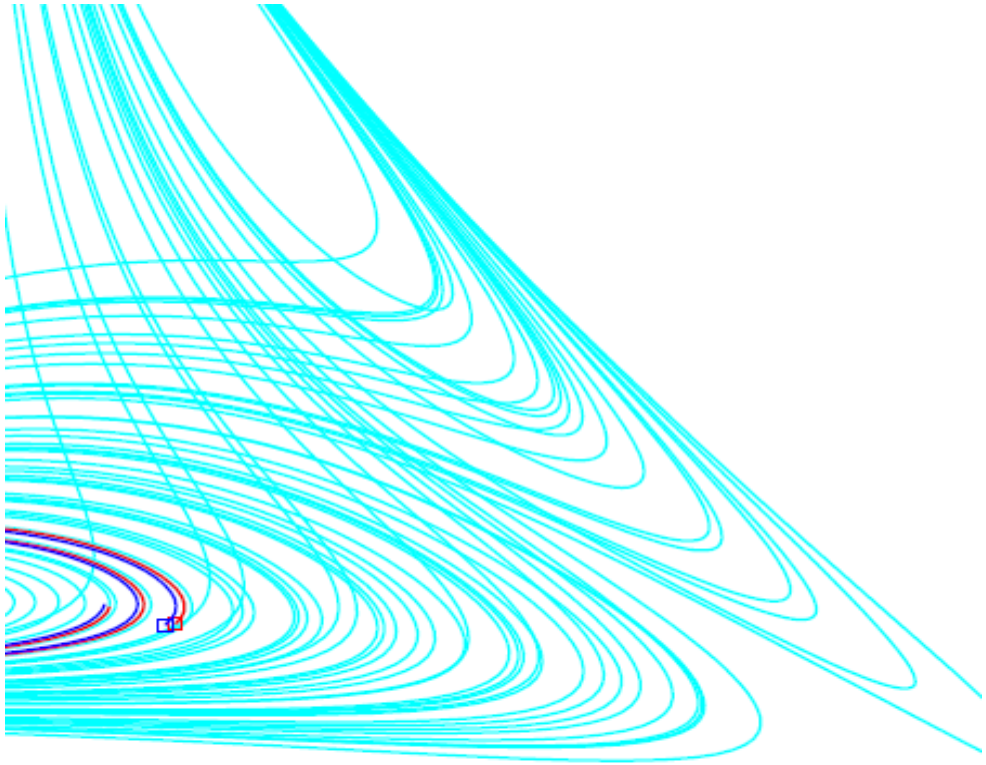


Returning to questions of what we would do if we had a perfect model:



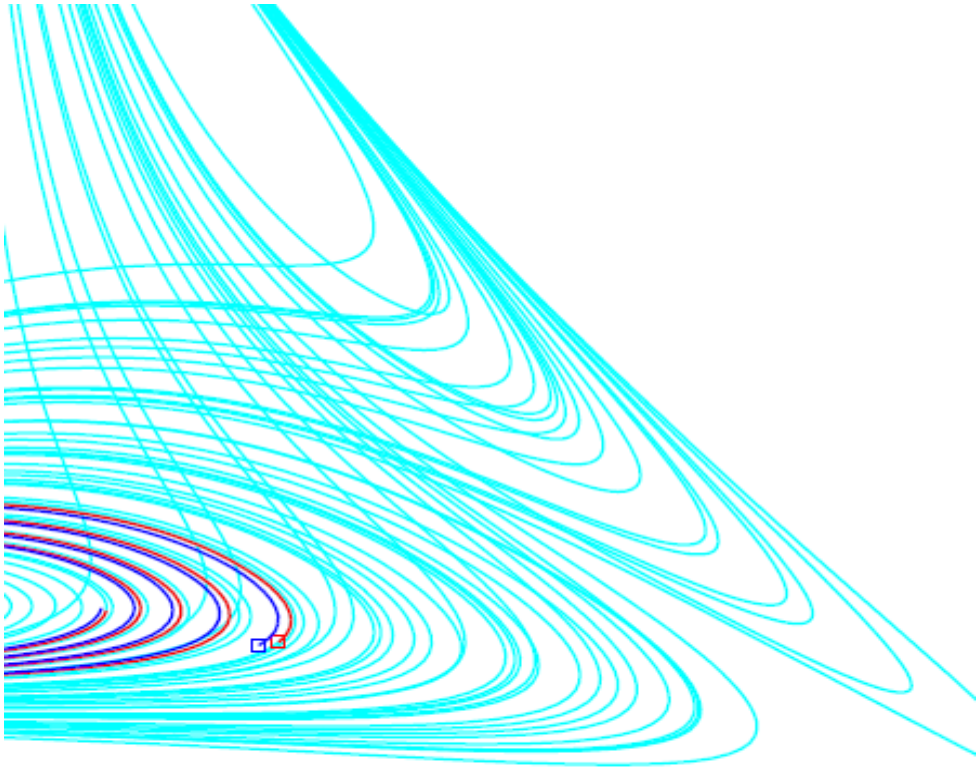
Images courtesy of Dr. Kevin Judd

Nonlinearity – Initial Value Sensitivity in weather



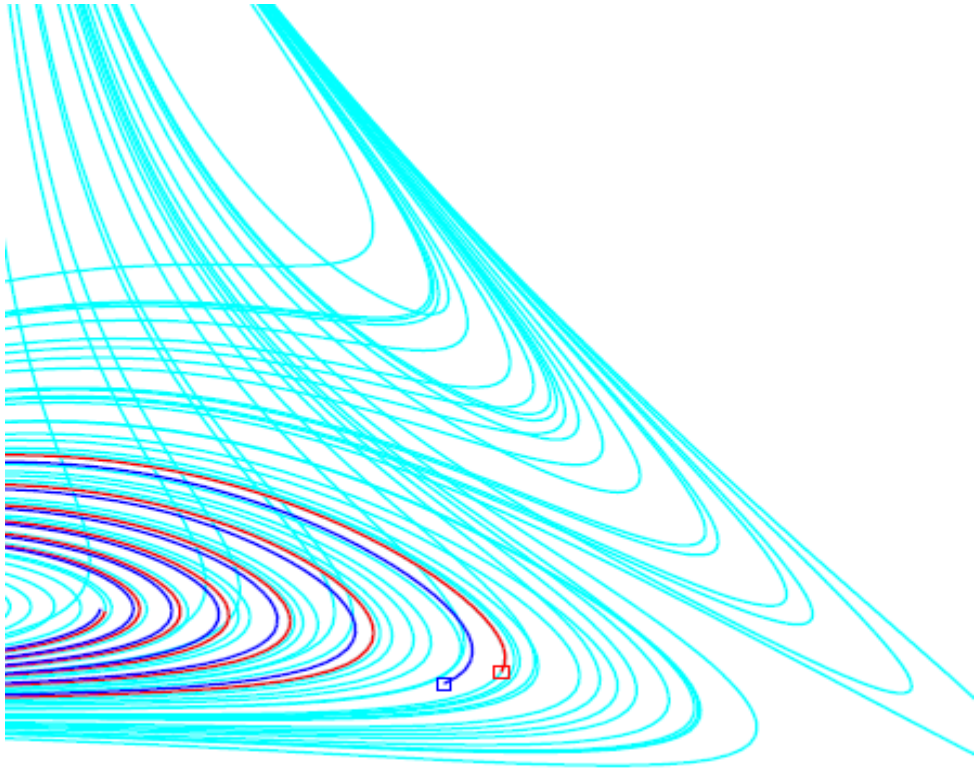
Images courtesy of Dr. Kevin Judd

Nonlinearity – Initial Value Sensitivity in weather



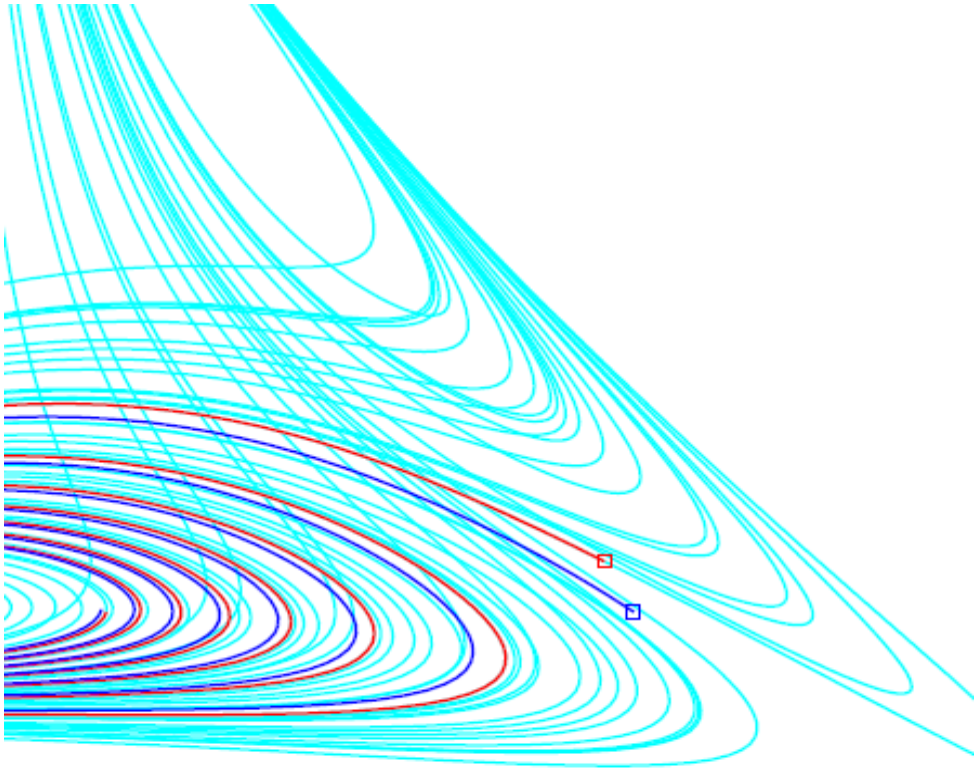
Images courtesy of Dr. Kevin Judd

Nonlinearity – Initial Value Sensitivity in weather



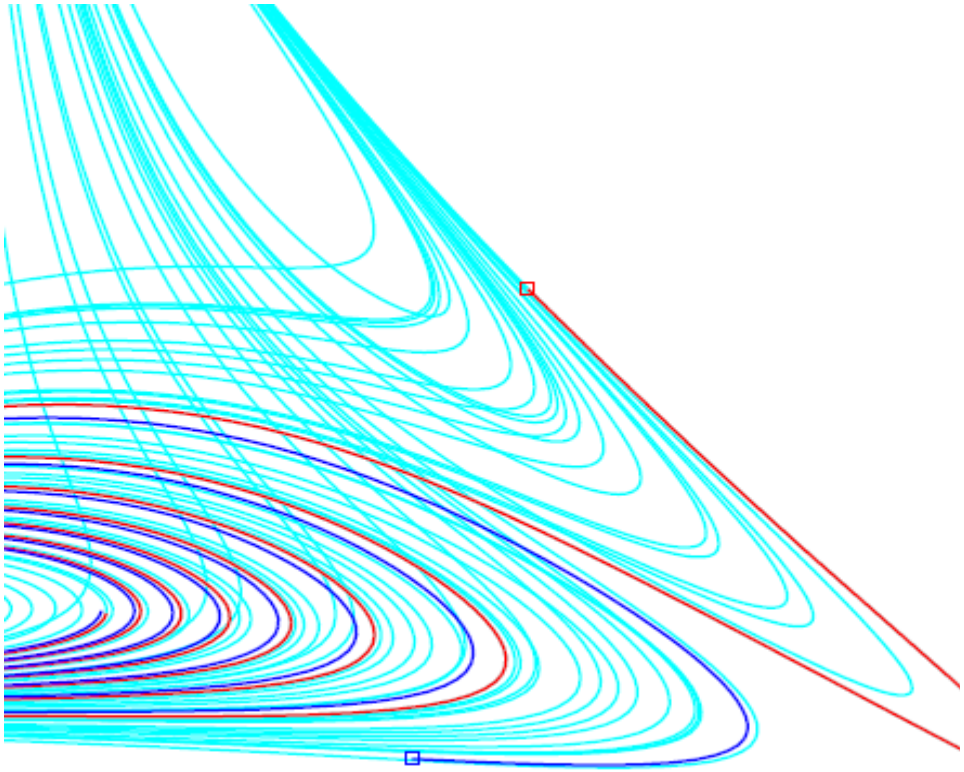
Images courtesy of Dr. Kevin Judd

Nonlinearity – Initial Value Sensitivity in weather



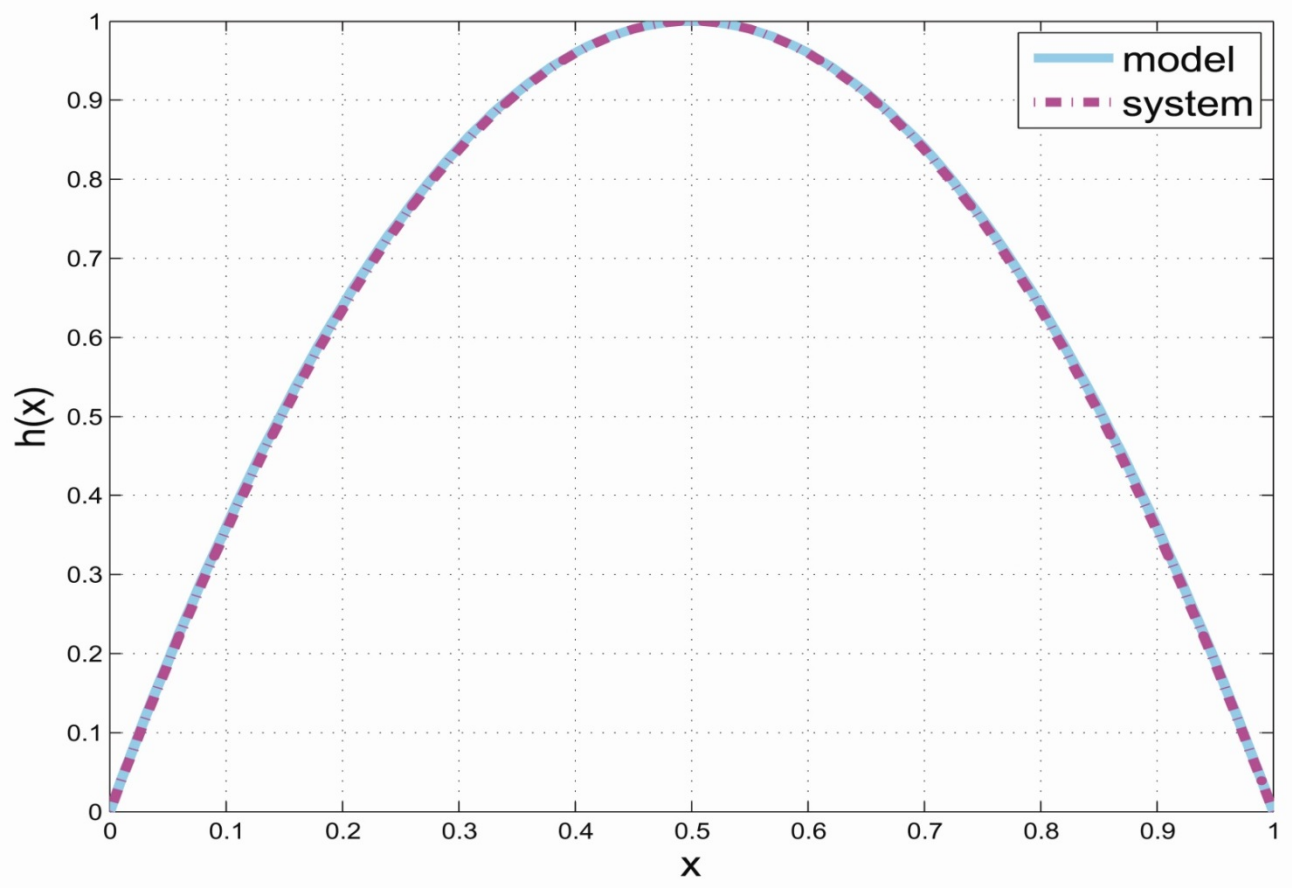
Images courtesy of Dr. Kevin Judd

Nonlinearity – Initial Value Sensitivity in weather



Images courtesy of Dr. Kevin Judd

The Logistic Map and the Hawkmoth Effect



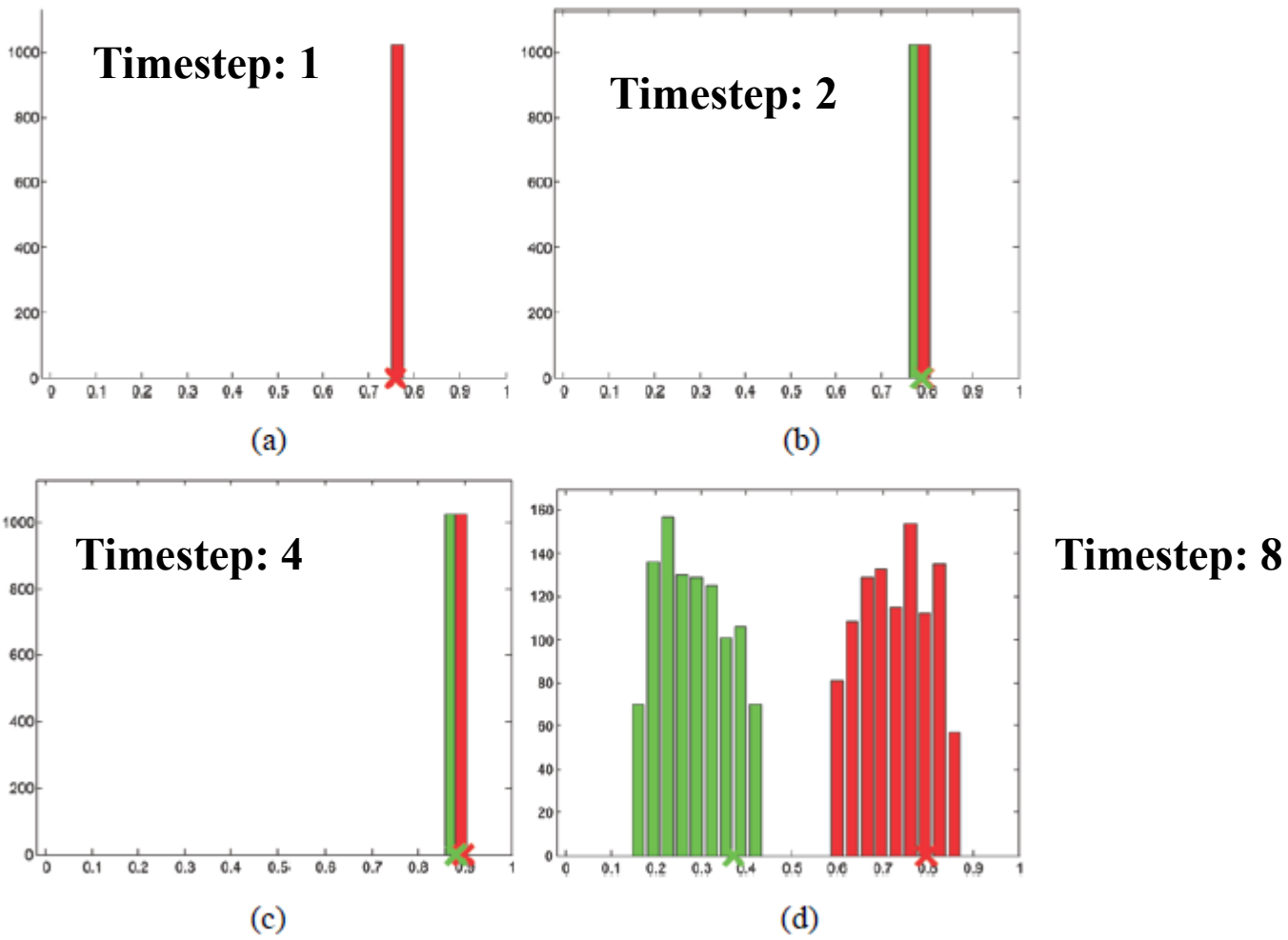
Model: $N_{t+1} = 4\,N_t(1- N_t)$

System: $N_{t+1} = 4N_t(1-N_t)\left[(1-\varepsilon)+\frac{4}{5}\varepsilon(N_t^{\,2}-N_t-1)\right]$

Laplace's Demon and Climate Change, Frigg et al., 2013

Nonlinearity –Sensitivity to Model Error

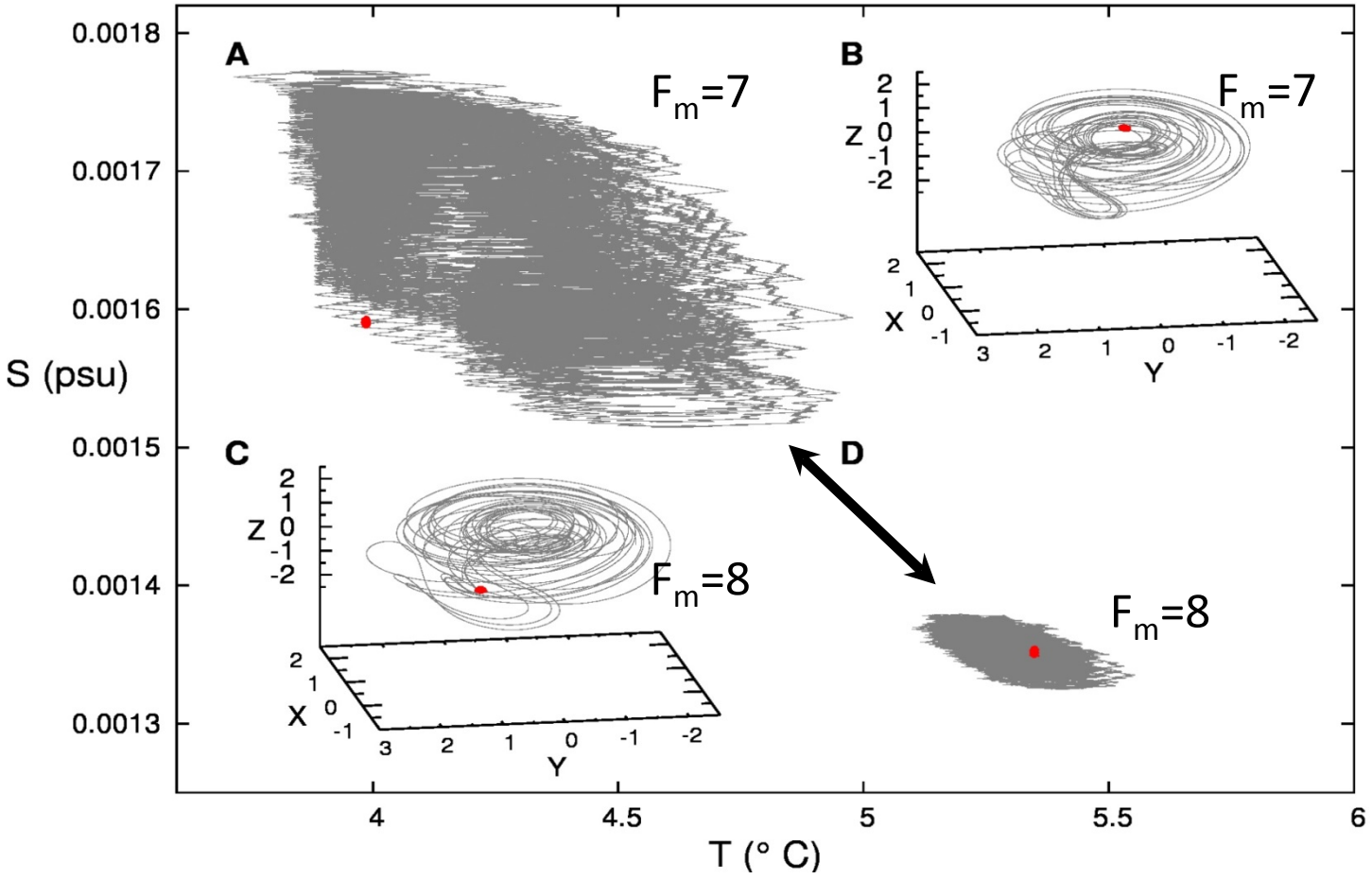
A Good Looking Model, Not A Good Forecasting System



Laplace's Demon and Climate Change, Frigg et al., 2013

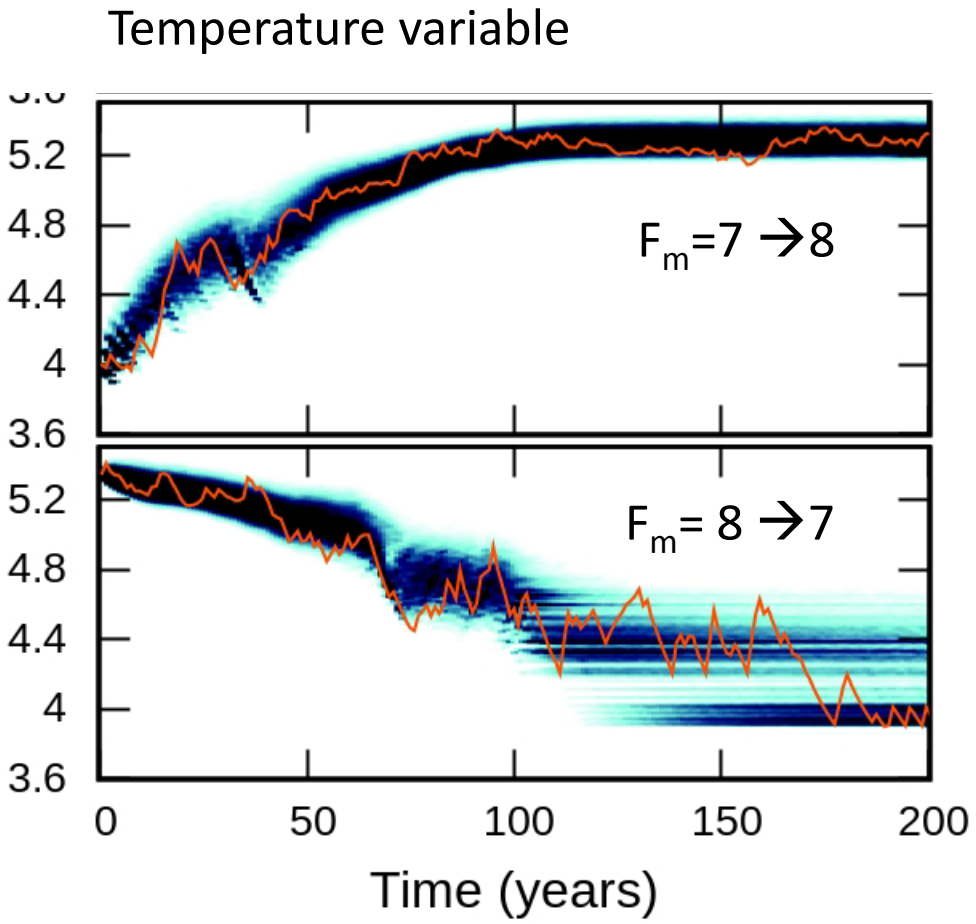
A Nonlinear System Experiment Which Parallels Climate Change

Moving From One attractor To Another



Daron and Stainforth, Env.Res.Lett., 2013

Initial Value Uncertainty and Climate Prediction

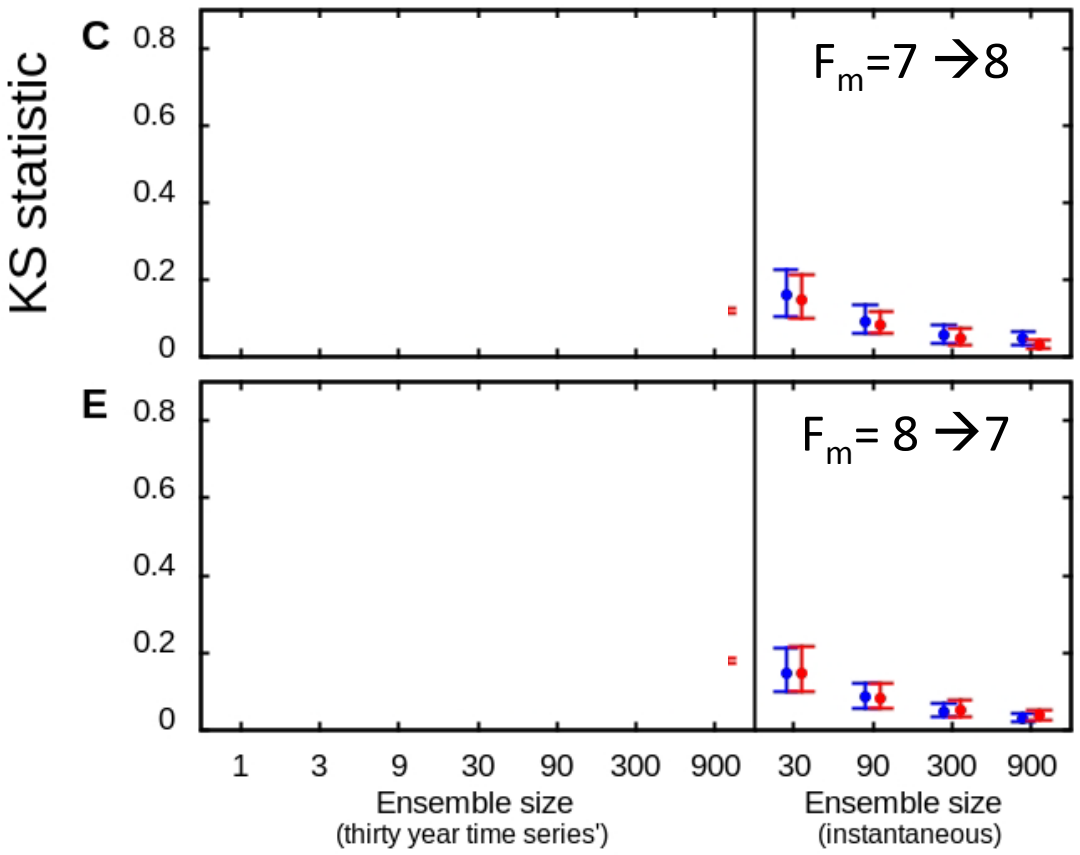


Frequency distributions from a 10,000 member initial-condition ensemble initiated from a single locale on the attractor.

Daron and Stainforth, Env.Res.Lett., 2013

How Big an Ensemble Do we Need?
Instantaneous Distributions

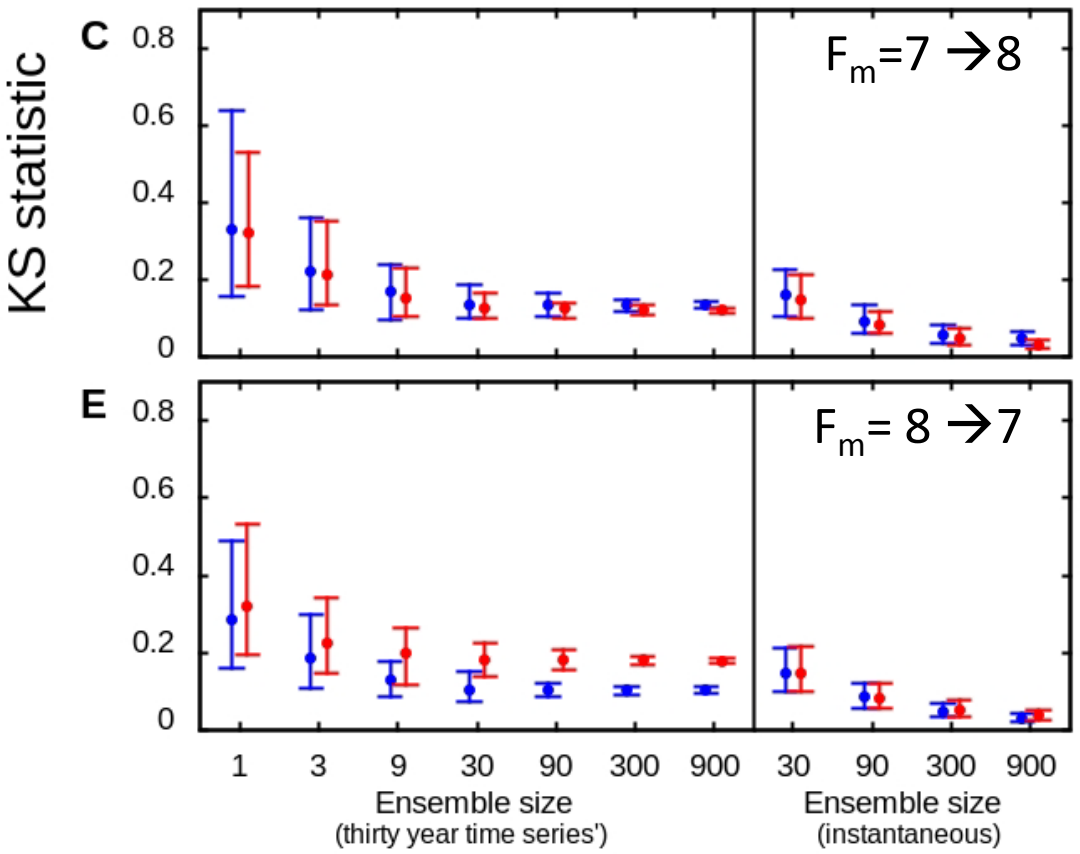
T variable



Daron and Stainforth, Env.Res.Lett., 2013

How Big an Ensemble Do we Need?
30 year Distributions About the Given Time Point

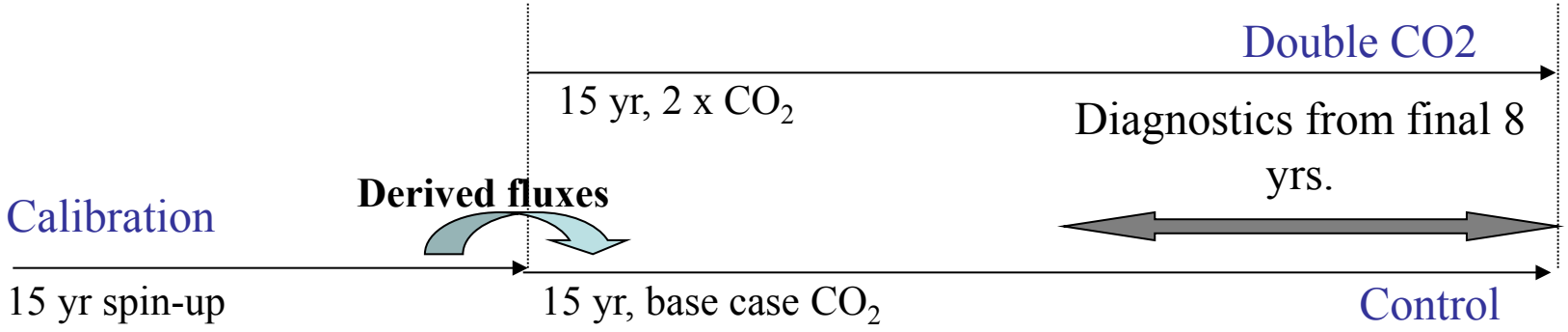
T variable

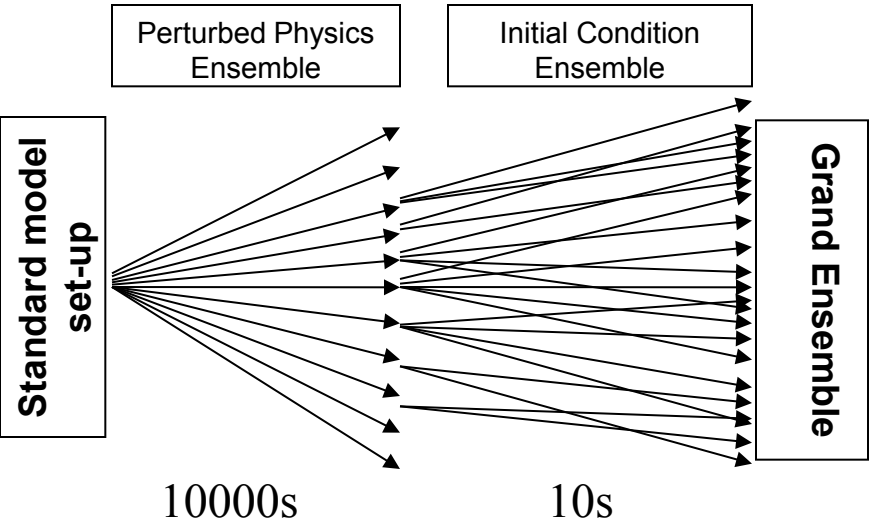


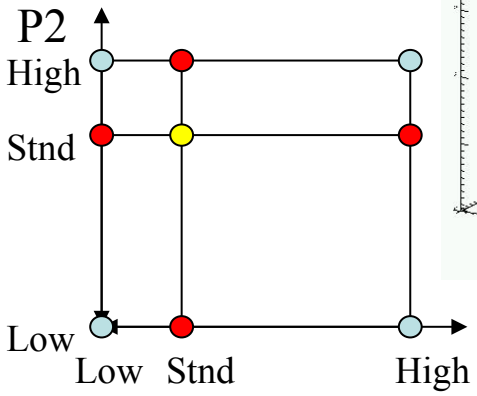
Daron and Stainforth, Env.Res.Lett., 2013

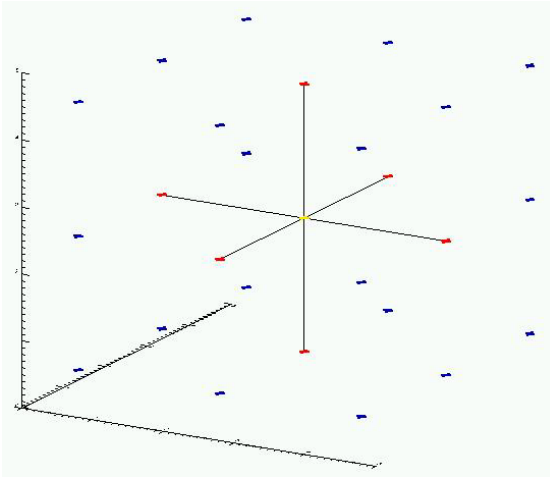
Climate*prediction*.net: The Slab Model Experiment

- Unified Model with thermodynamic ocean. (HadSM3)

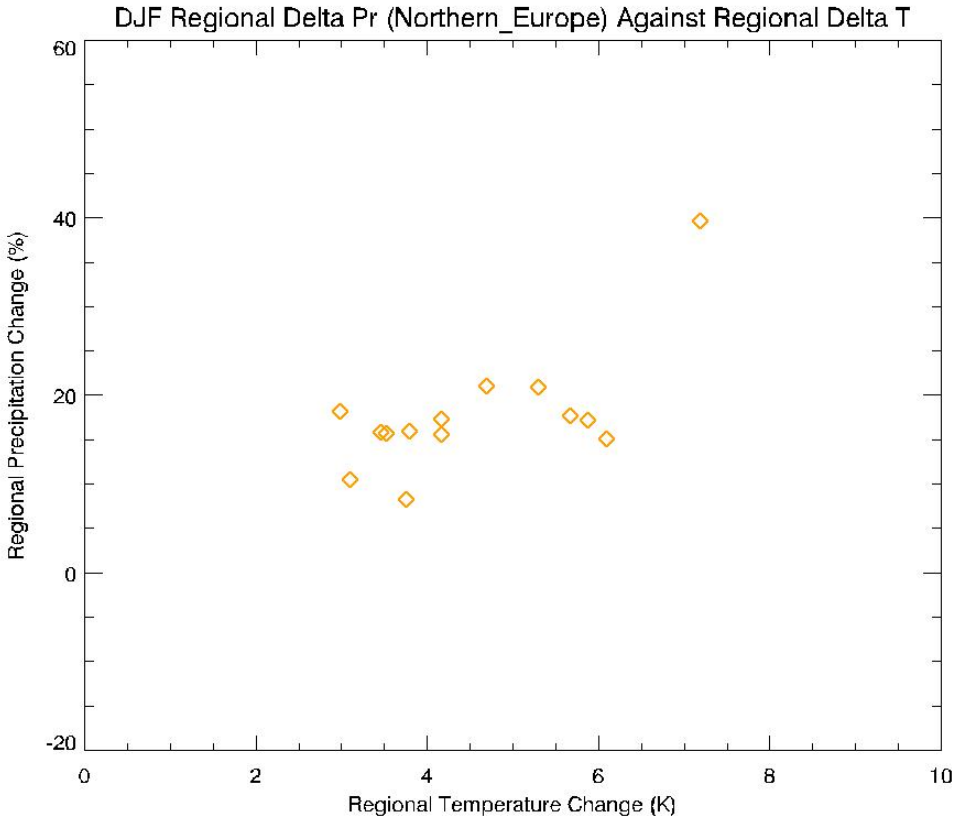


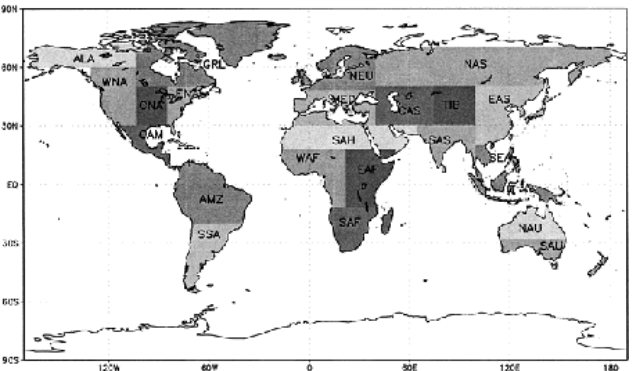


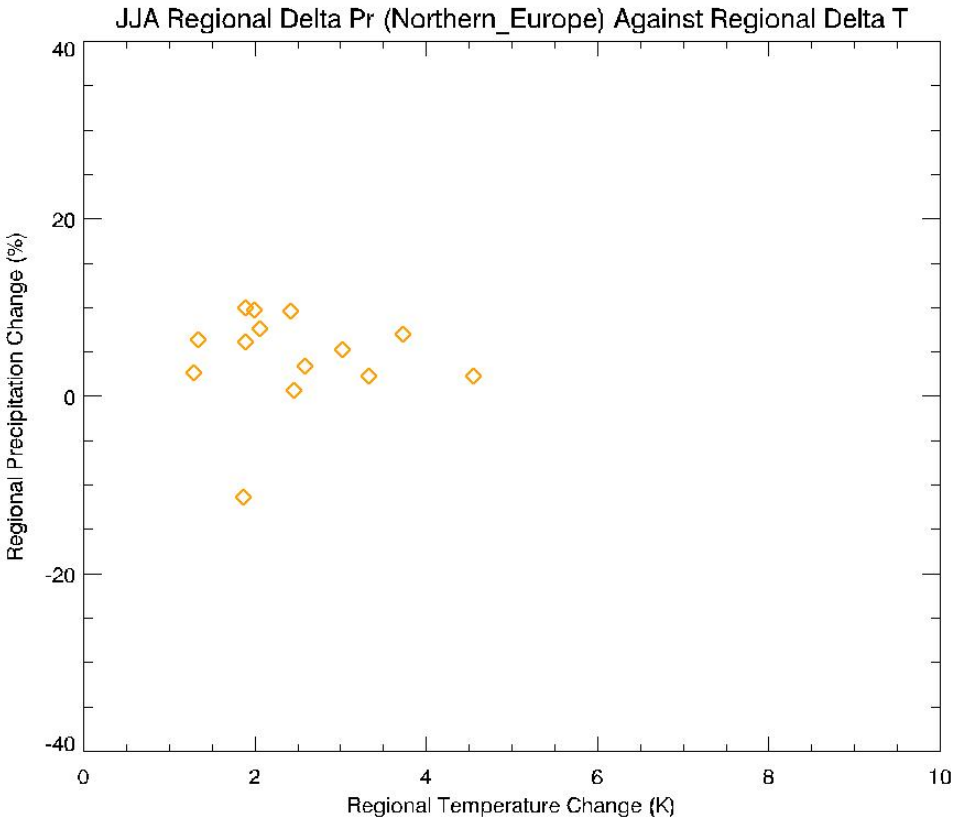




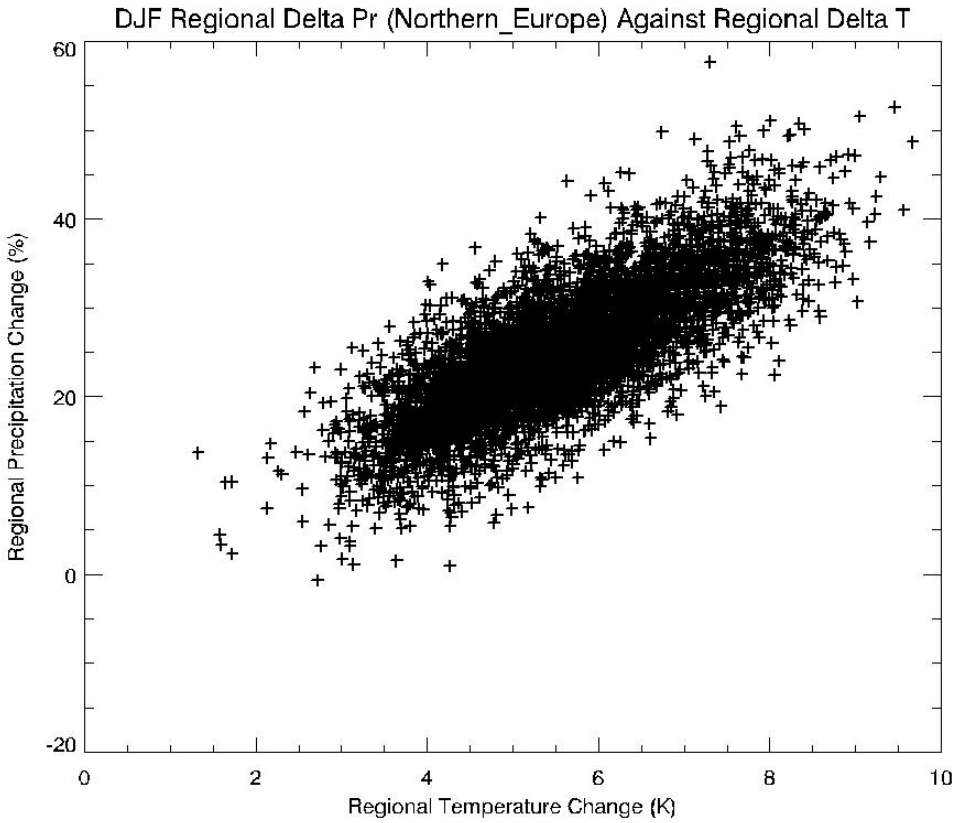
Multi-Model Regional Distributions



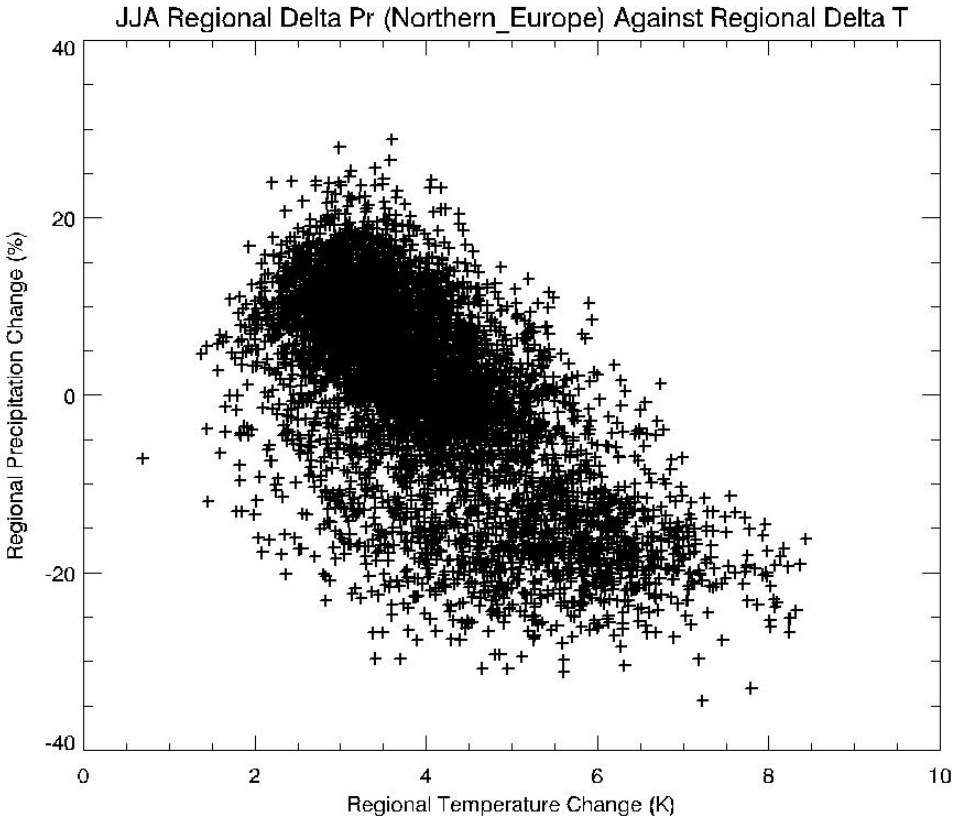




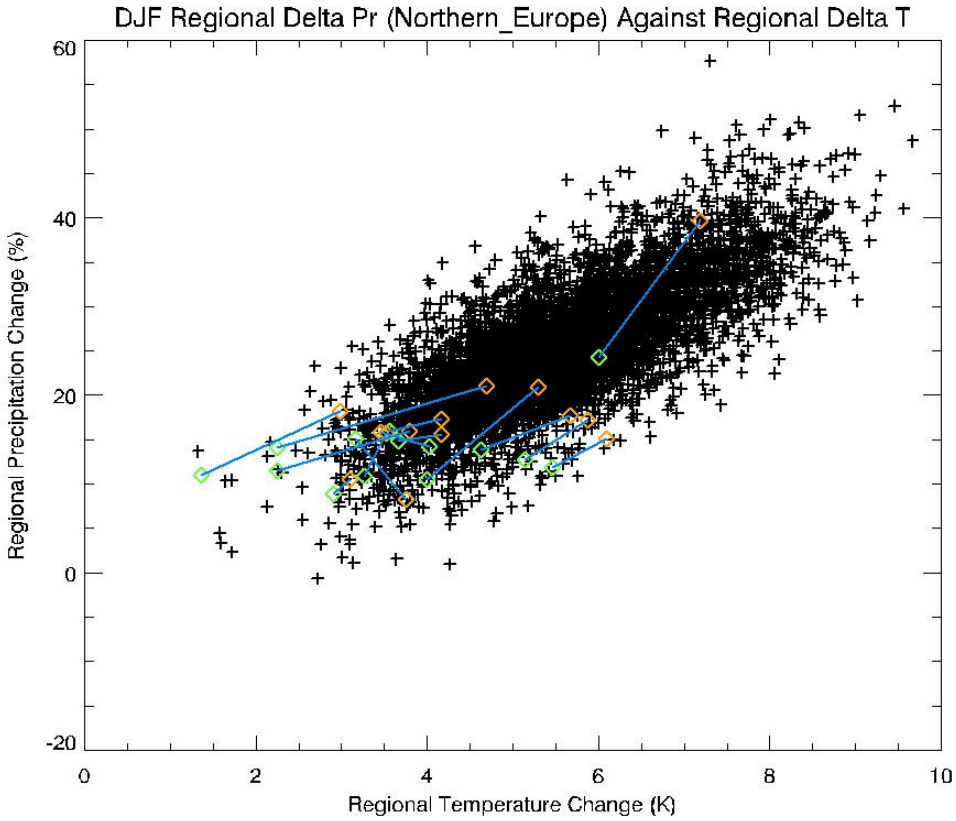
Regional Distributions



- 20,000 simulations
- 6203 model versions with points representing average over initial condition ensembles.

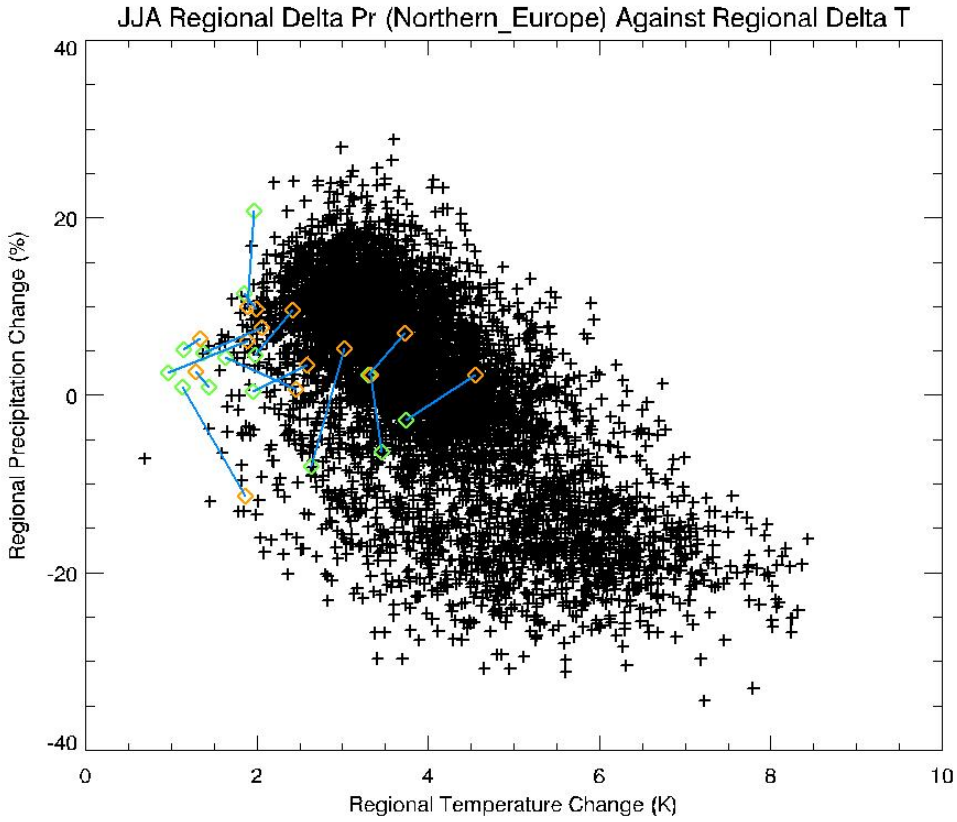


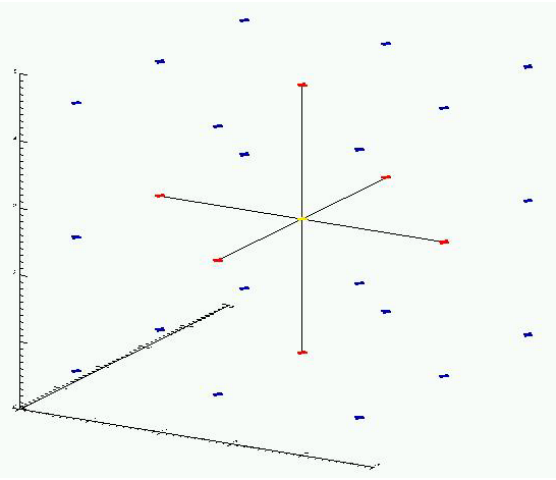
Regional Distributions

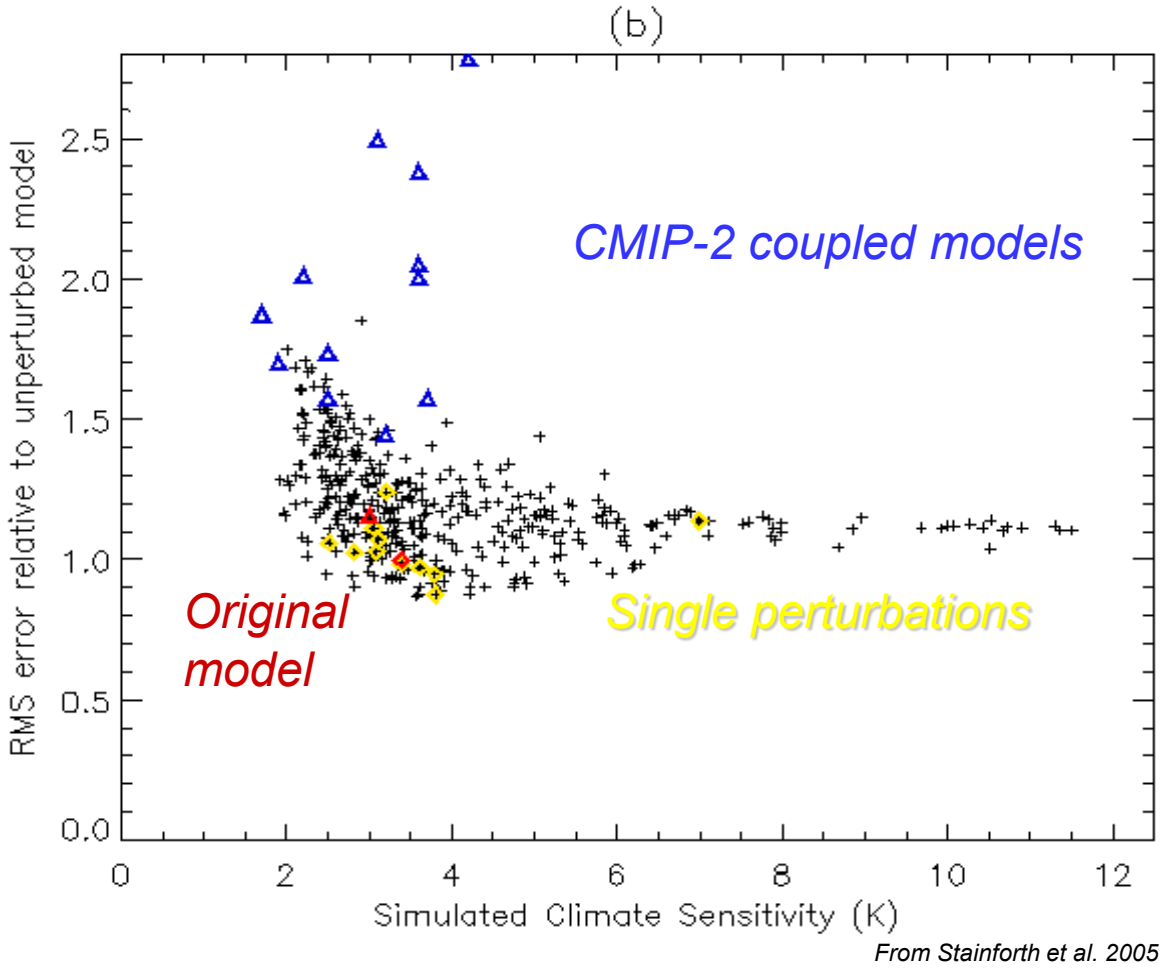


Challenge 1: Lack of Independence

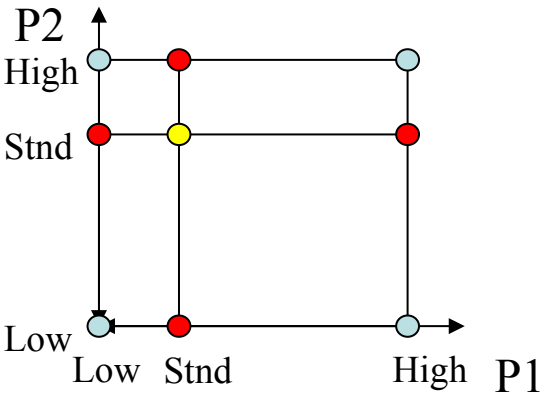
- The model versions are highly dependent on each other.
- High density of points does not relate to greater probability.



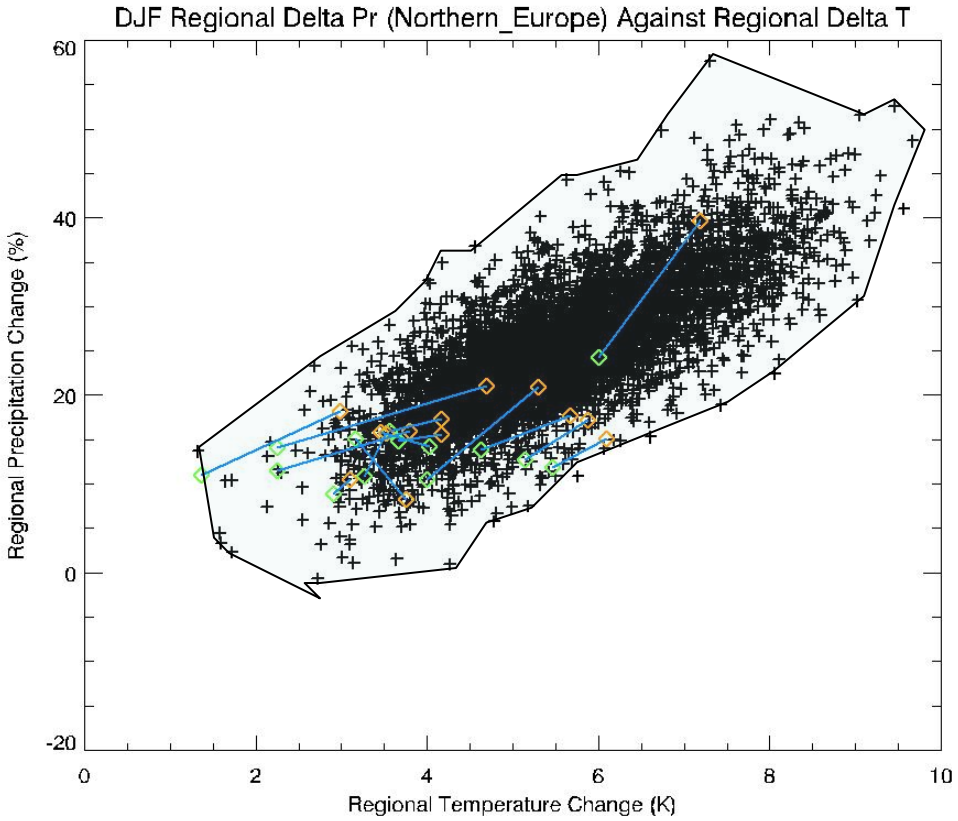




From Stainforth et al. 2005

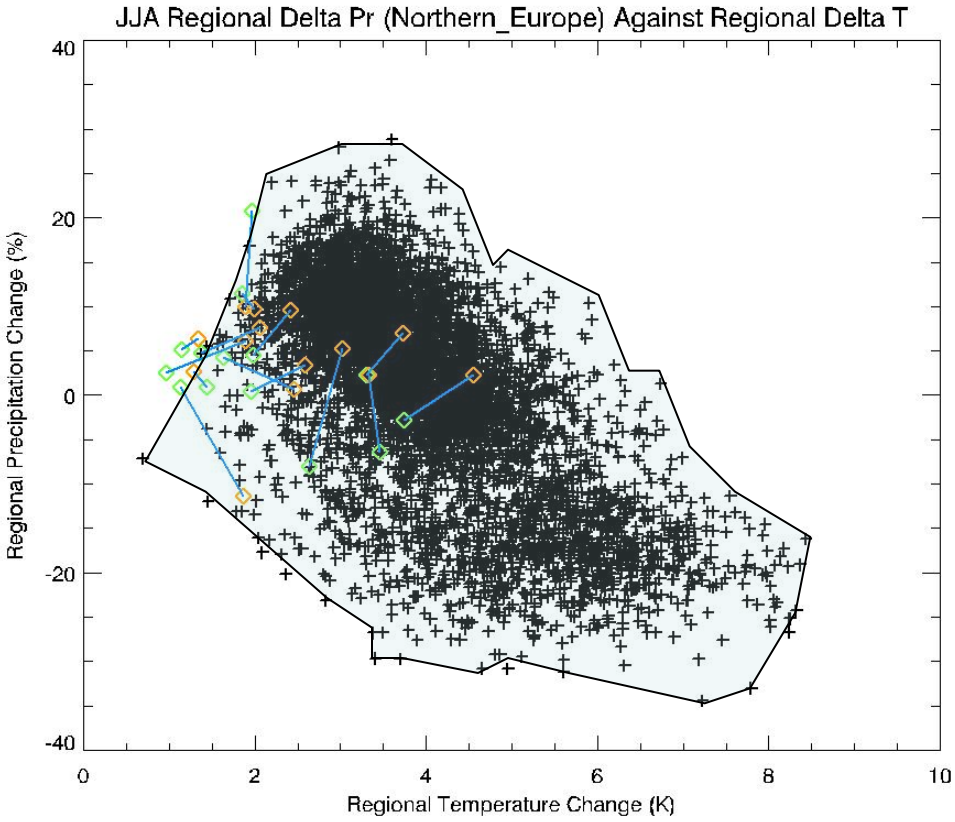


Regional Distributions

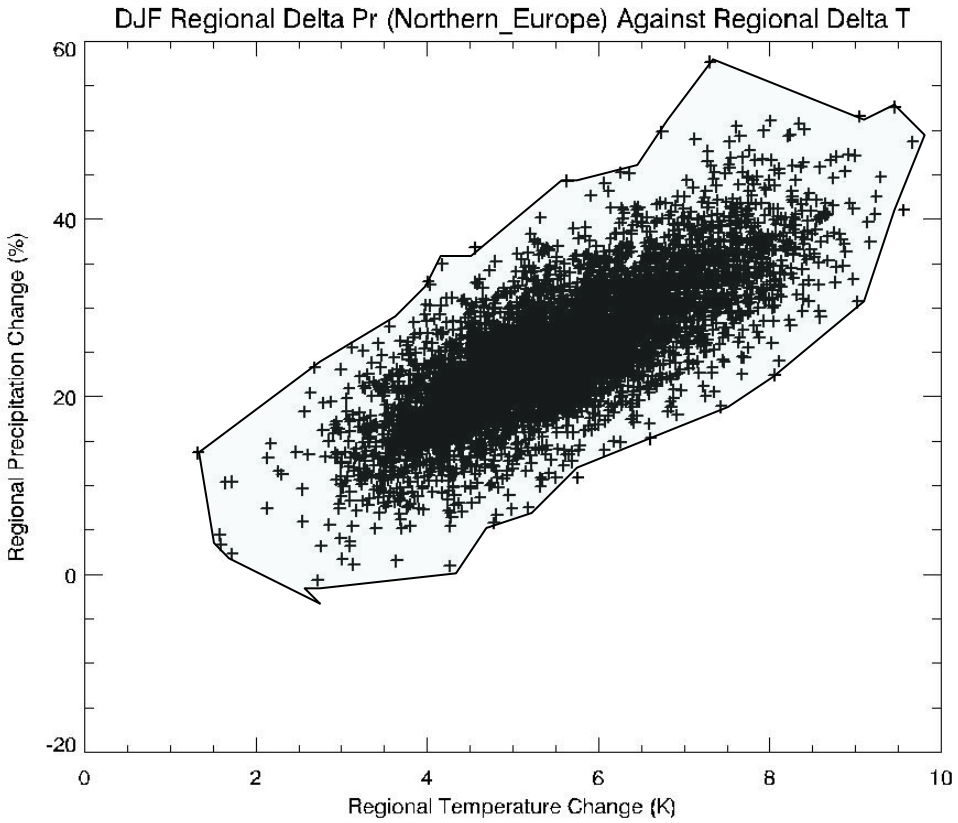


Challenge 1: Lack of Independence

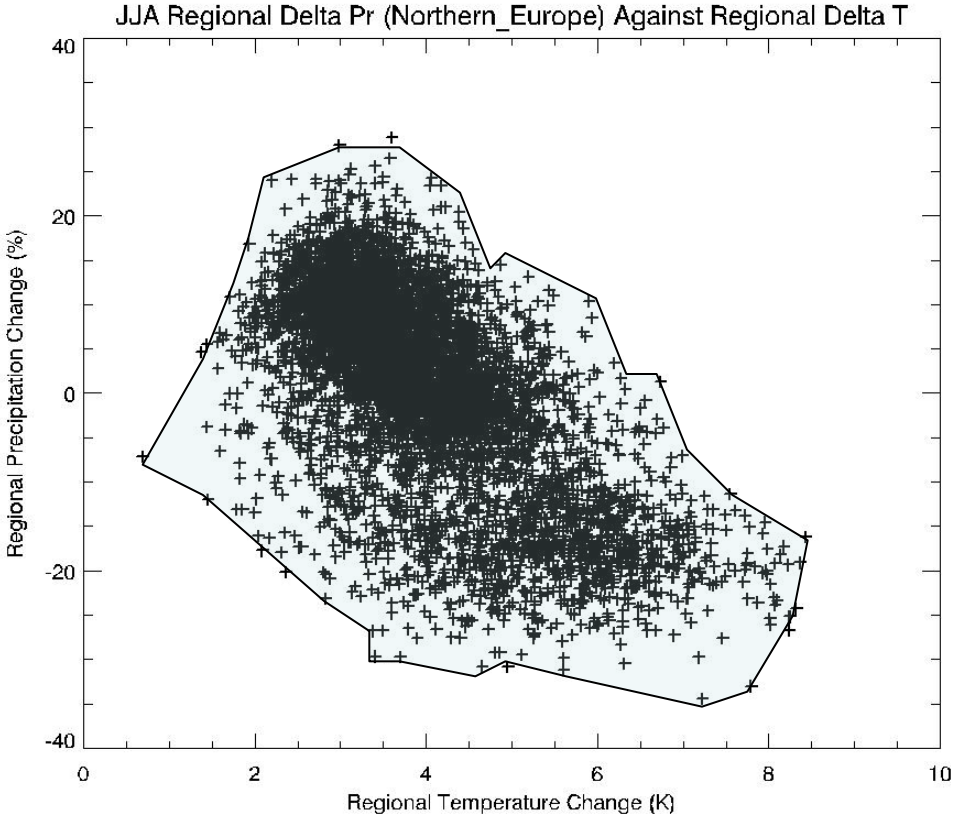
- The model versions are highly dependent on each other.
- High density of points does not relate to greater probability.



To the extent that any simulations are a plausible future, they all are:



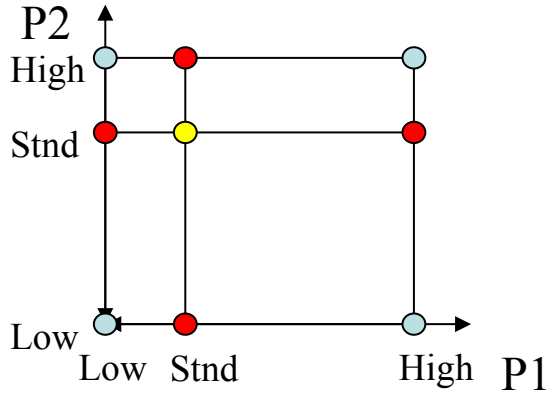
"Domain of possibility"
"Non-discountable envelope"[2]
"Lower bound on the maximum range of uncertainty"[1]

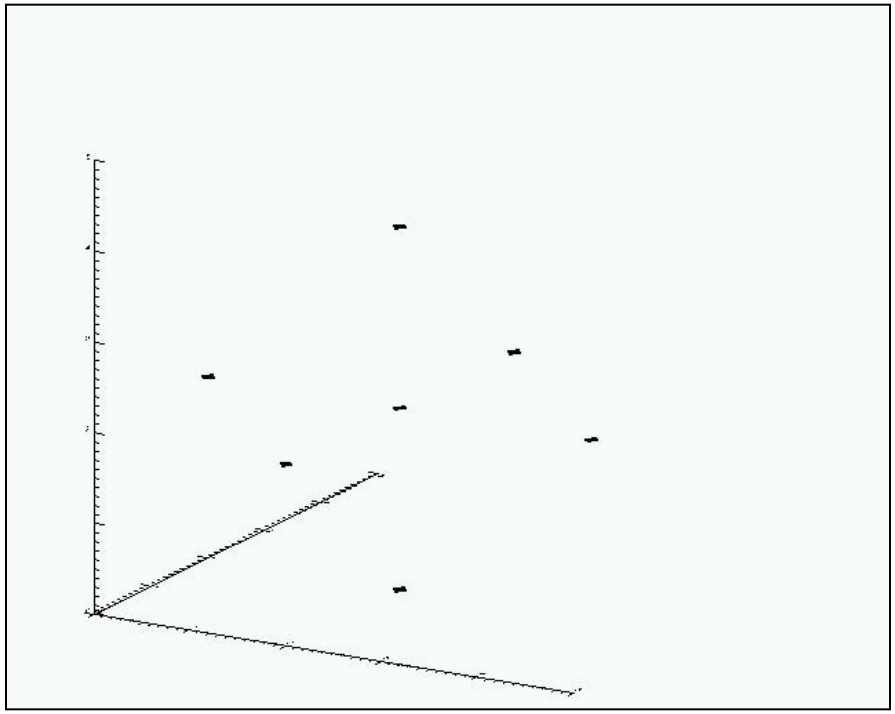


1. Stainforth et al., **Confidence, uncertainty and decision-support relevance in climate predictions**. Phil Trans Roy Soc365 (1857), 2145 (2007).
2. Stainforth et al. **Issues in the interpretation of climate model ensembles to inform decisions.** Phil Trans Roy Soc. 365 (1857), 2163 (2007).

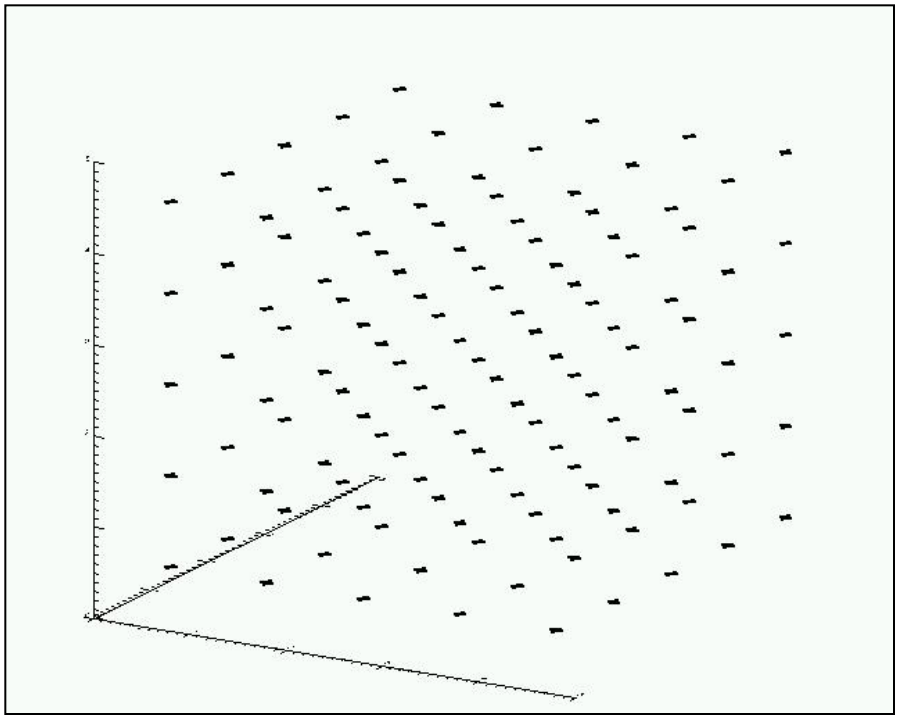
Lack of Independence, Emulation and Sampling Design

- What about filling in parameter space with an emulator?



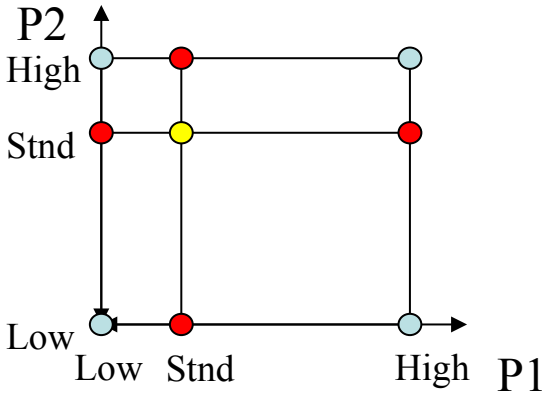


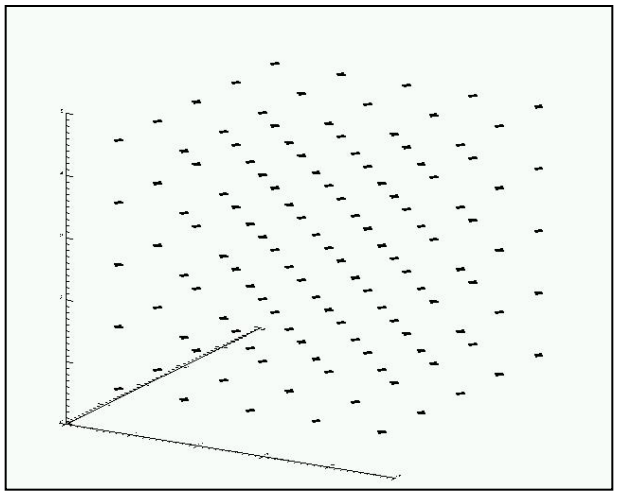
Emulate



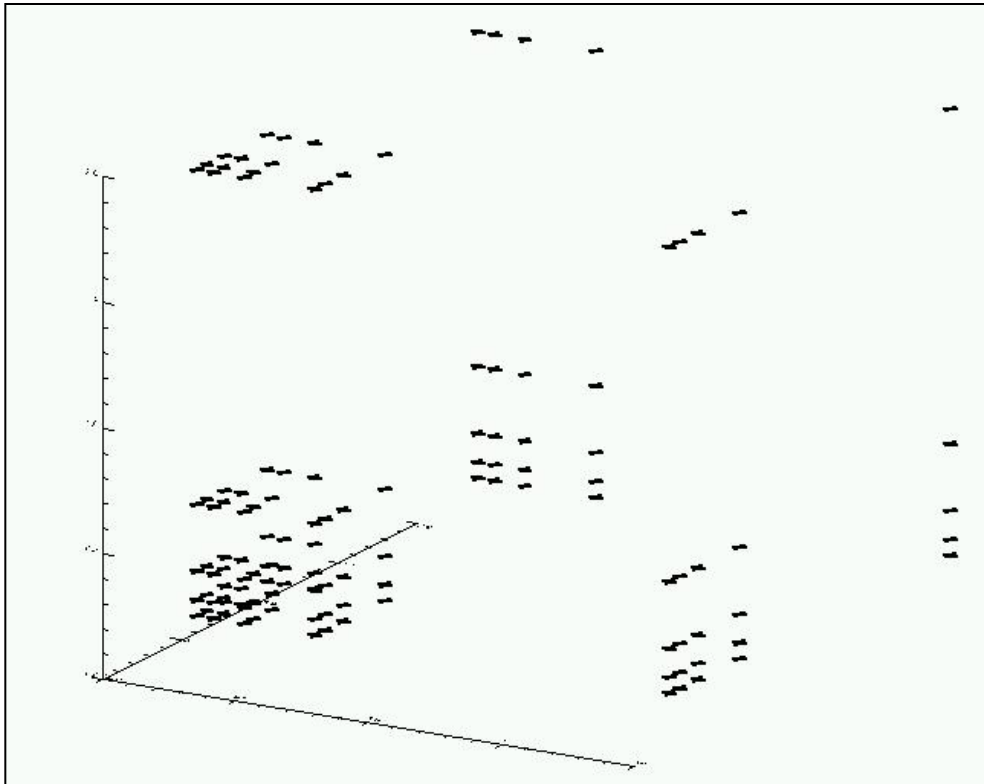
Lack of Independence Revisited

- What about filling in parameter space with an emulator?
- Unfortunately there is no objective prior there.
- **Even the shape of parameter space (and of model space if one could define it) is arbitrary.**



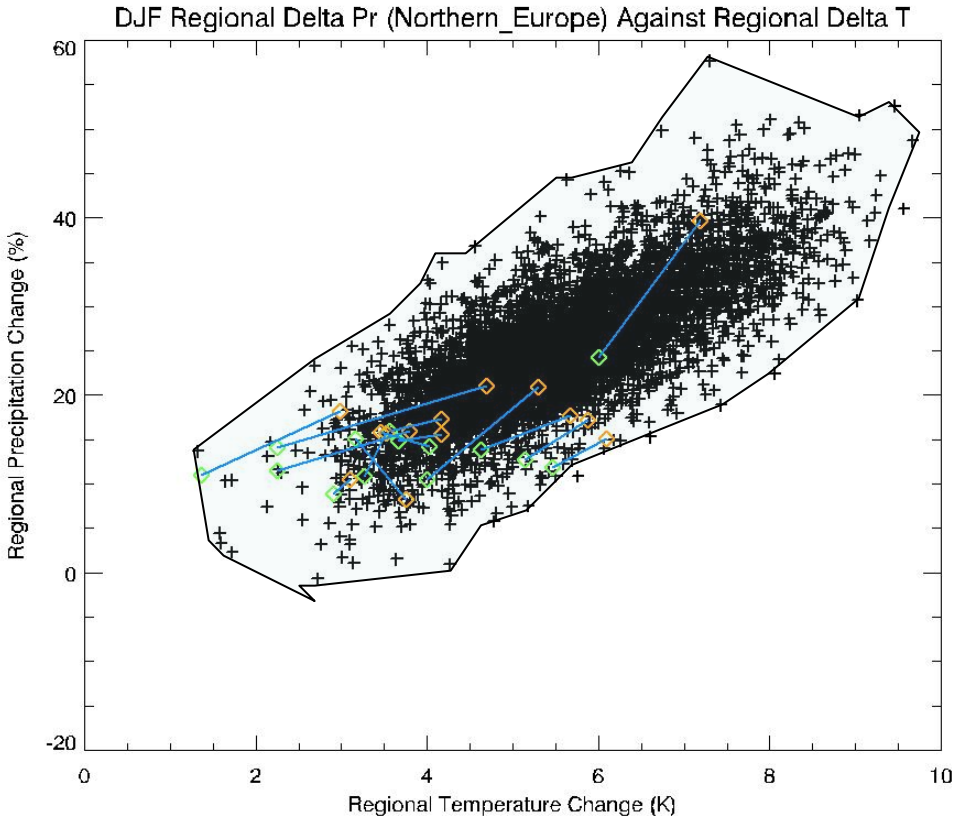


Choice of parameter definition



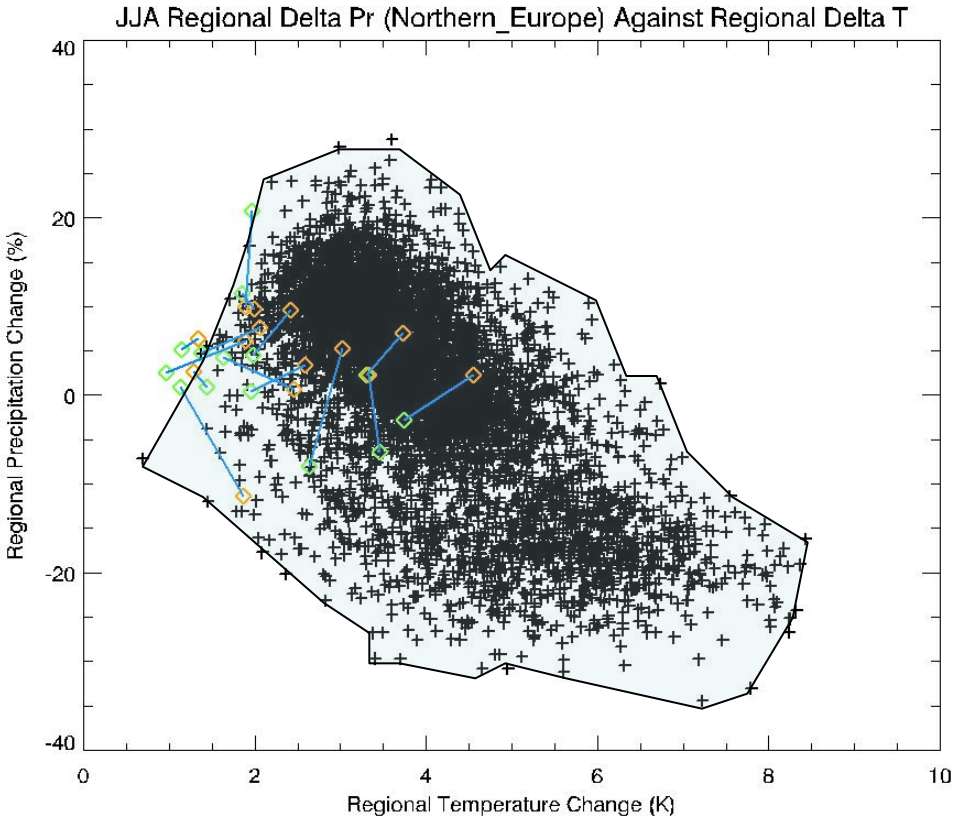
- **How do these parameters relate to reality?**
- **What's the meaning of "cloud ice fall rate" in a 200km square grid box?**

Regional Distributions

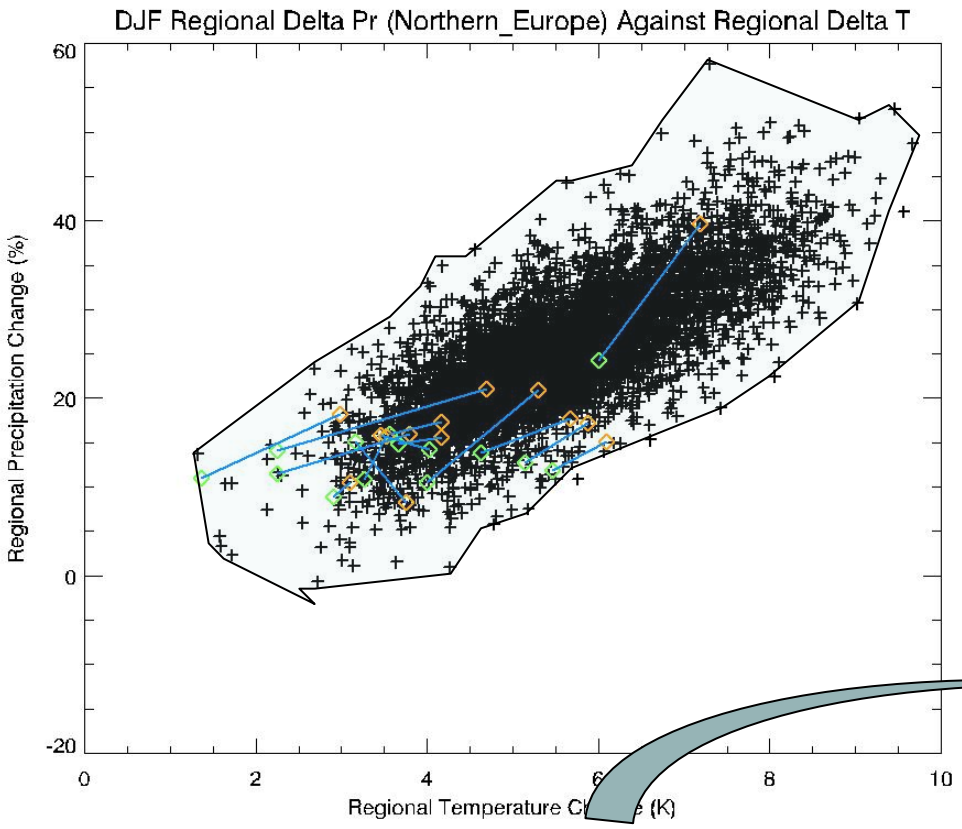


Challenge 2: In-Sample Analysis:

- Out-of-sample data can not be obtained in the future.
- Once published, further analysis becomes biased.
- Suggestion: Community agrees to hold back sample for future verification.



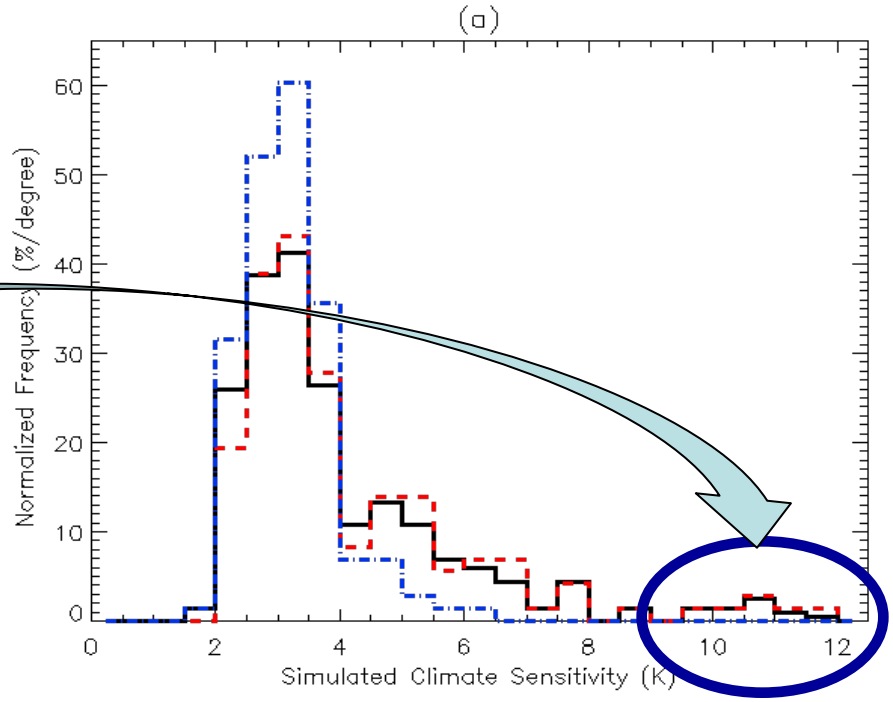
A Conflict of Physics and Statistics



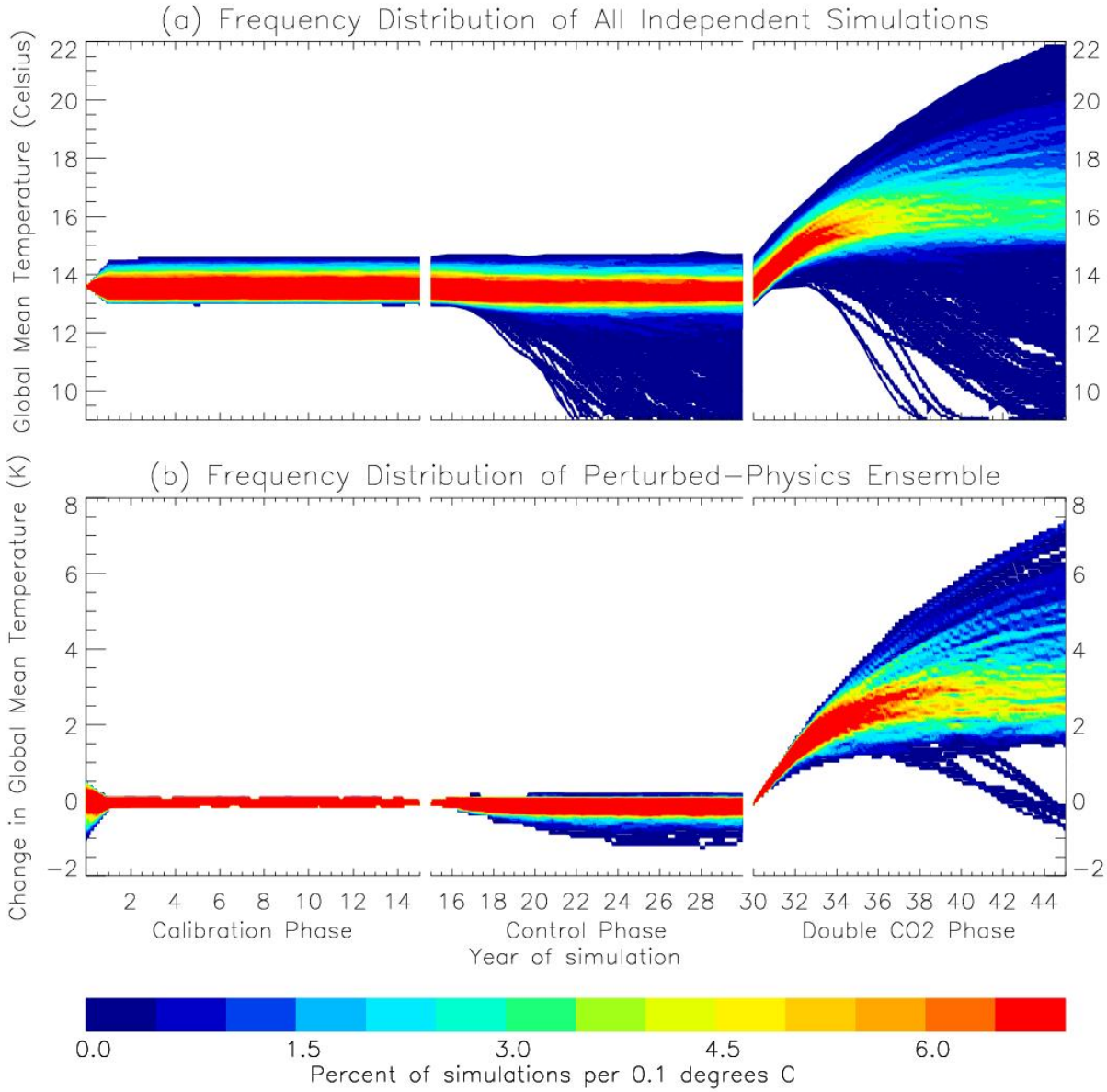
Low Entrainment Coefficent:
Rodwell & Palmer, 2007.
Joshi et al., 2011

Challenge 2: In-Sample Analysis:

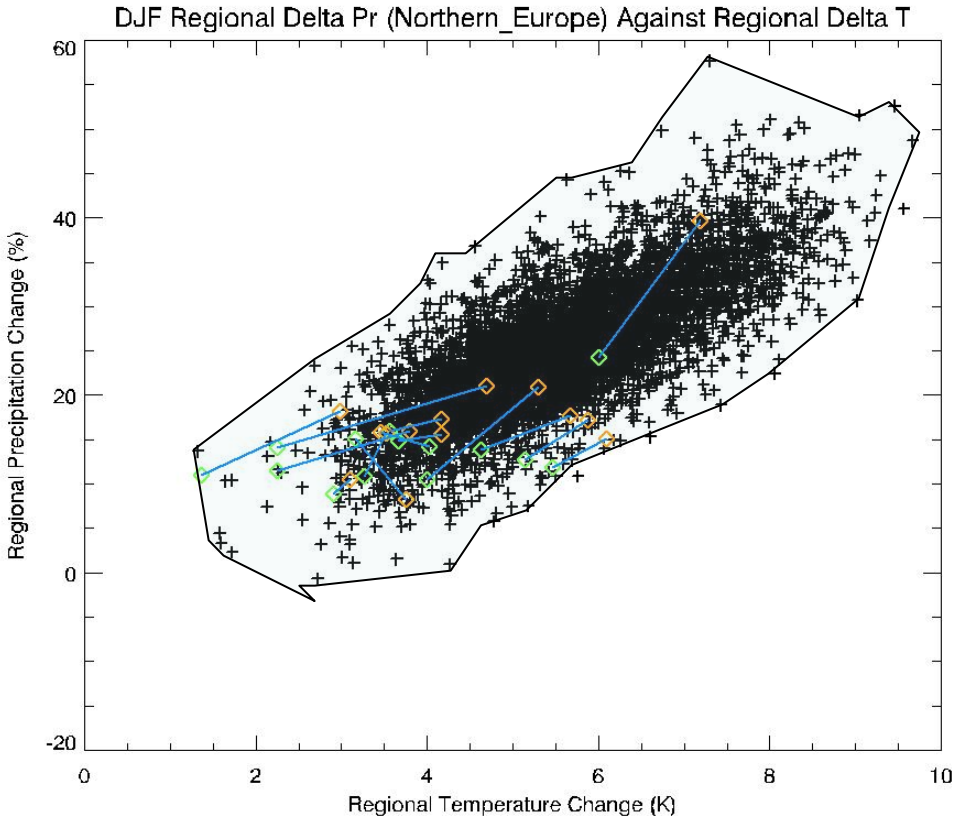
- Out-of-sample data can not be obtained in the future.
- Once published, further analysis becomes biased.
- Suggestion: Community agrees to hold back sample for future verification.



Stainforth et al., 2005

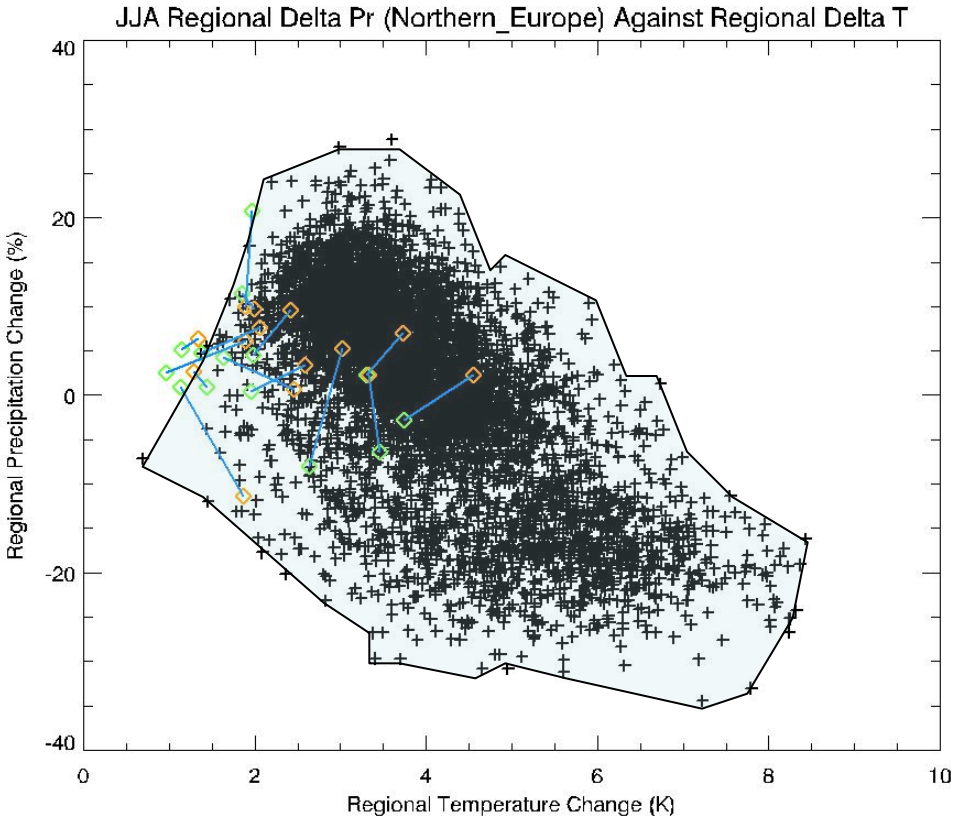

(a) Frequency Distribution of All Independent Simulations
Global Mean Temperature (Celsius)
22
20
18
16
14
12
10
(b) Frequency Distribution of Perturbed−Physics Ensemble
Change in Global Mean Temperature (K)
8
6
4
2
0
−2
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44
Calibration Phase
Control Phase
Double CO2 Phase
Year of simulation
0.0
1.5
3.0
4.5
6.0
Percent of simulations per 0.1 degrees C

Handling The In-Sample Problem
Don't Look at All your Data?

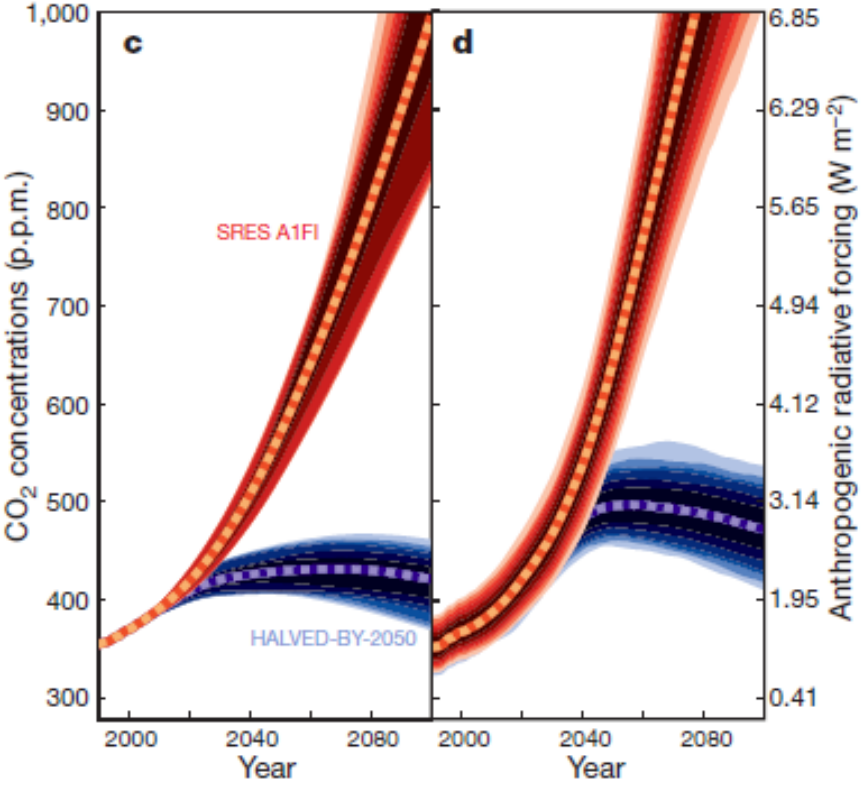


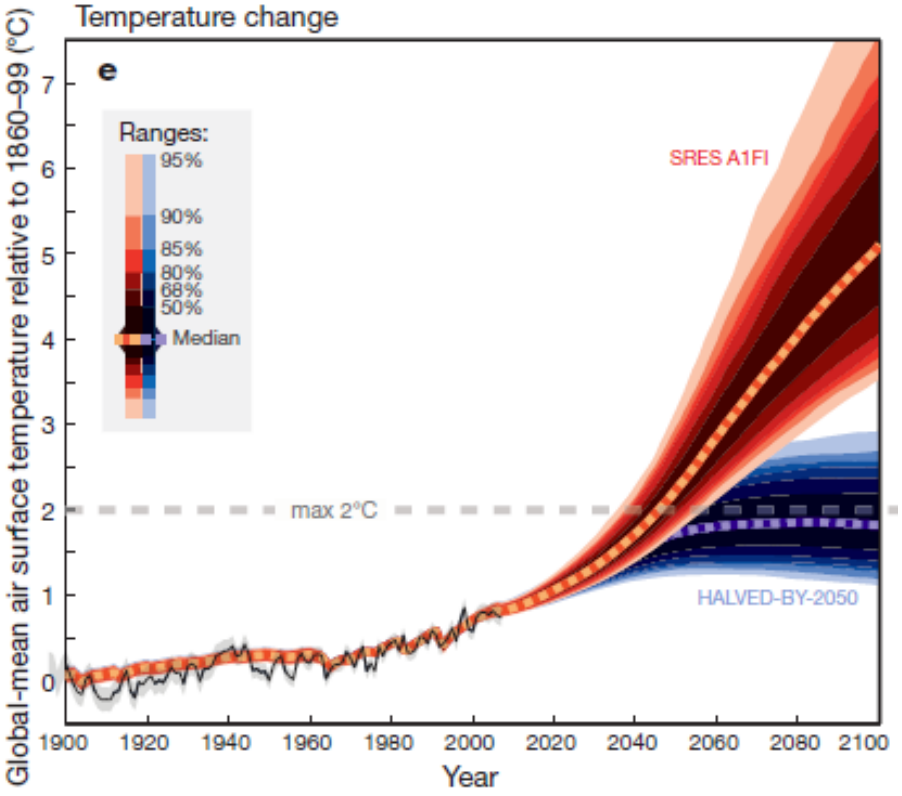
Challenge 2: In-Sample Analysis:

- Out-of-sample data can not be obtained in the future.
- Once published, further analysis becomes biased.
- Suggestion: Community agrees to hold back sample for future verification.



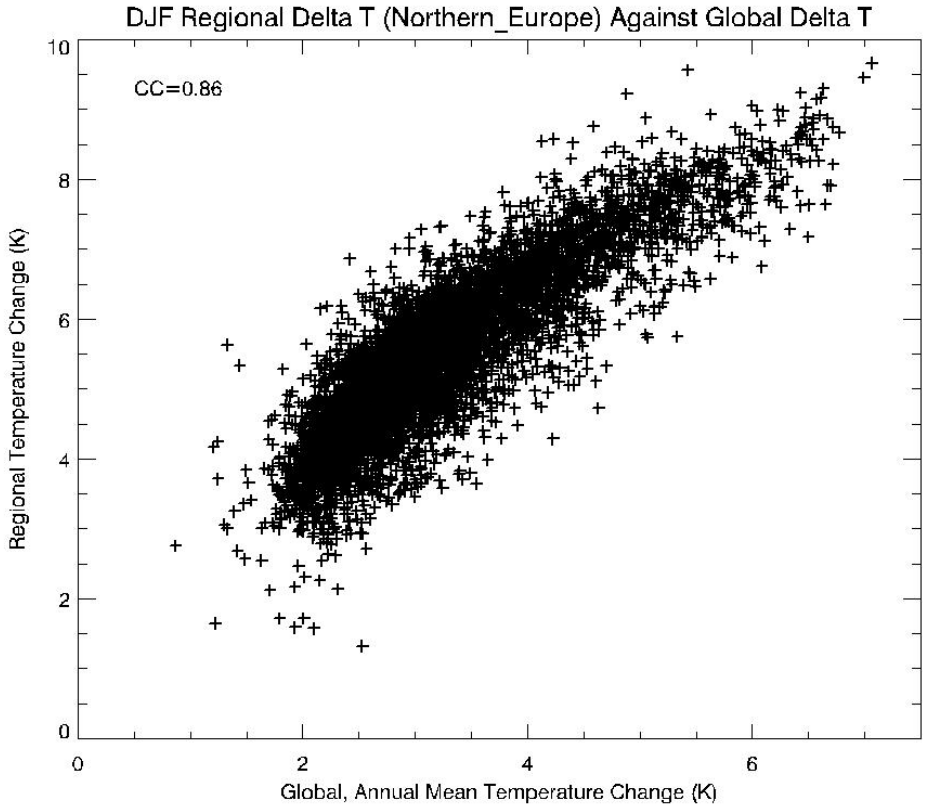
Maybe we have probabilities for global mean temperature?

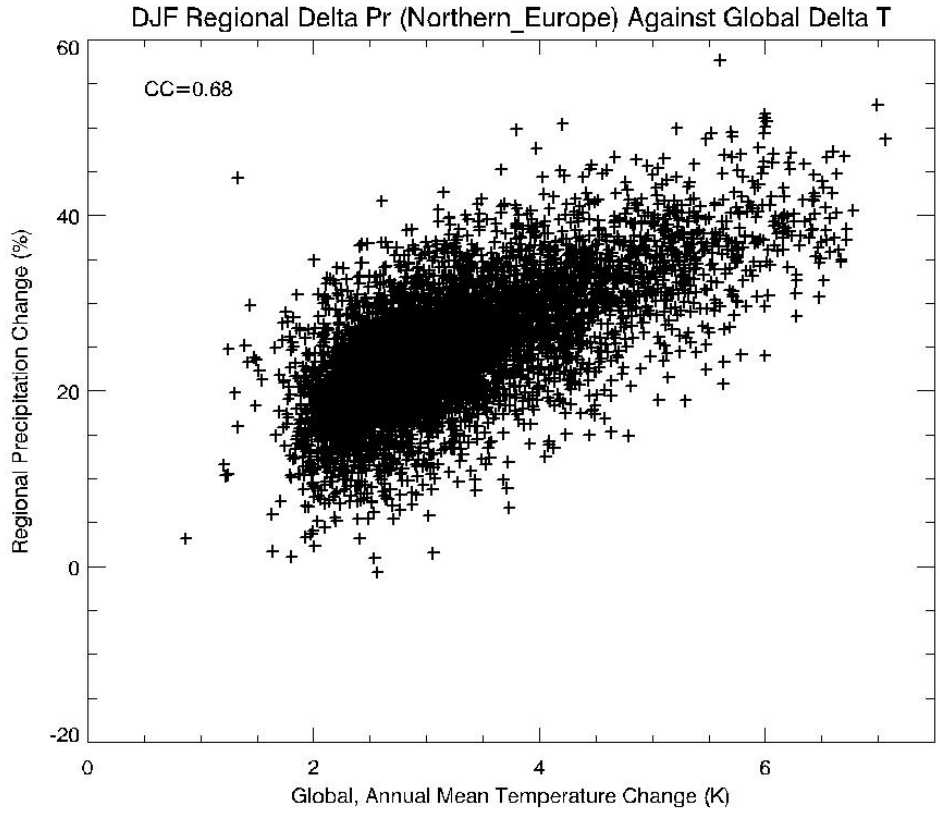




Meinshaussen et al., Nature, 2009

2 – Regional Change .vs. Global Temperature Change

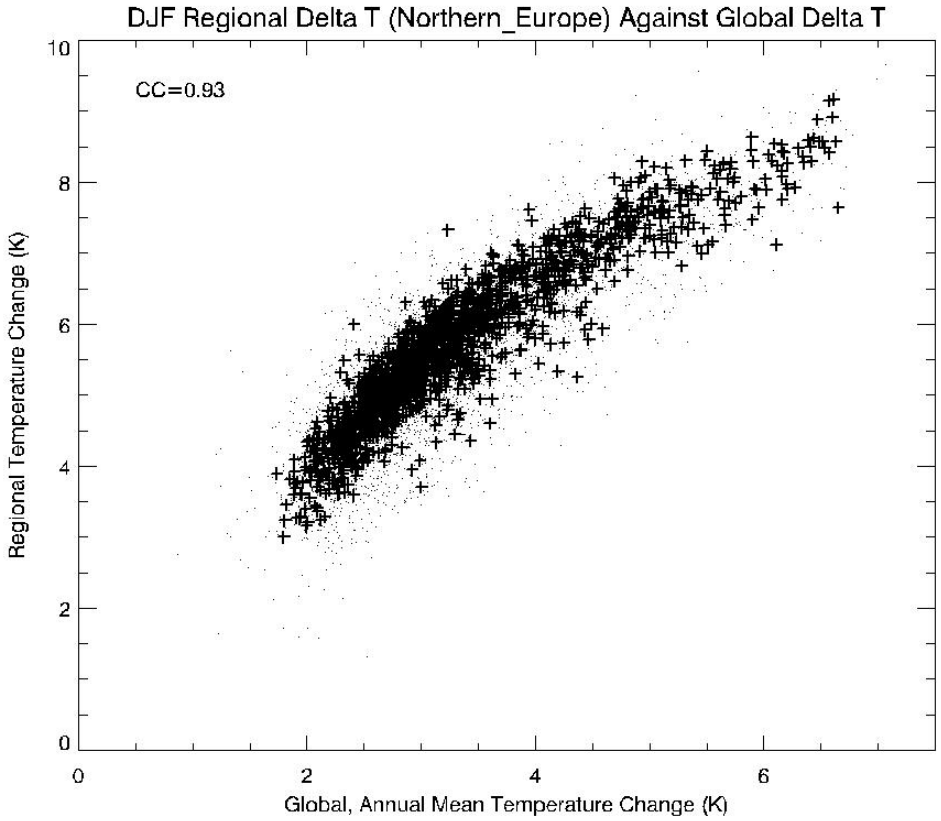


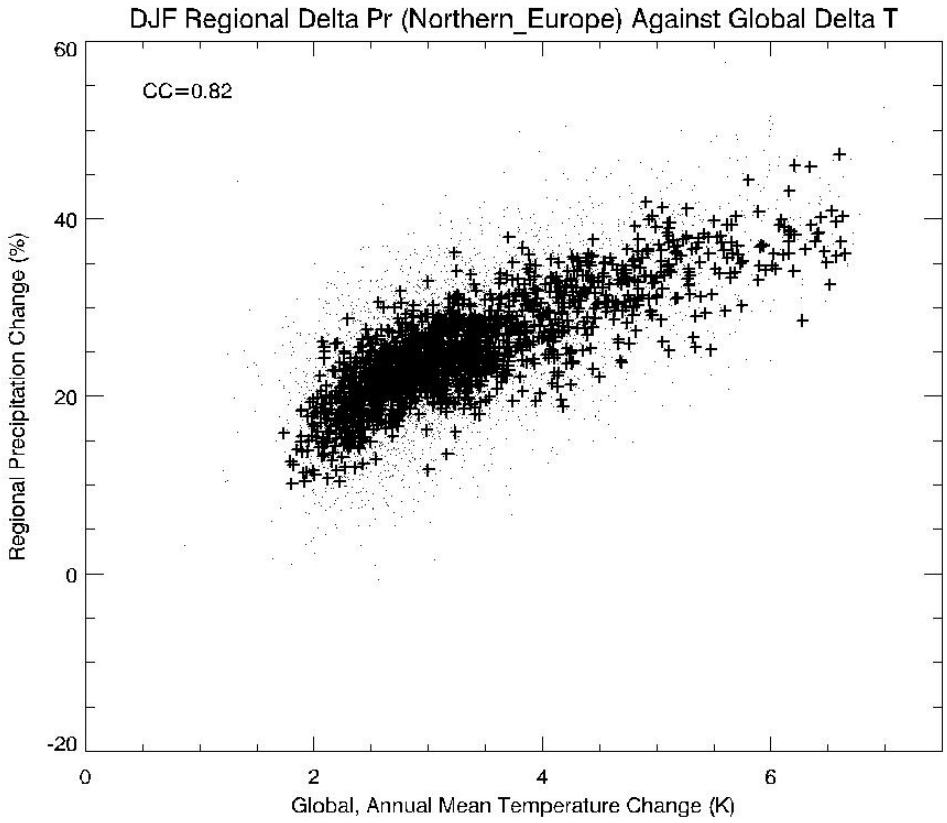


Ensemble Sizes

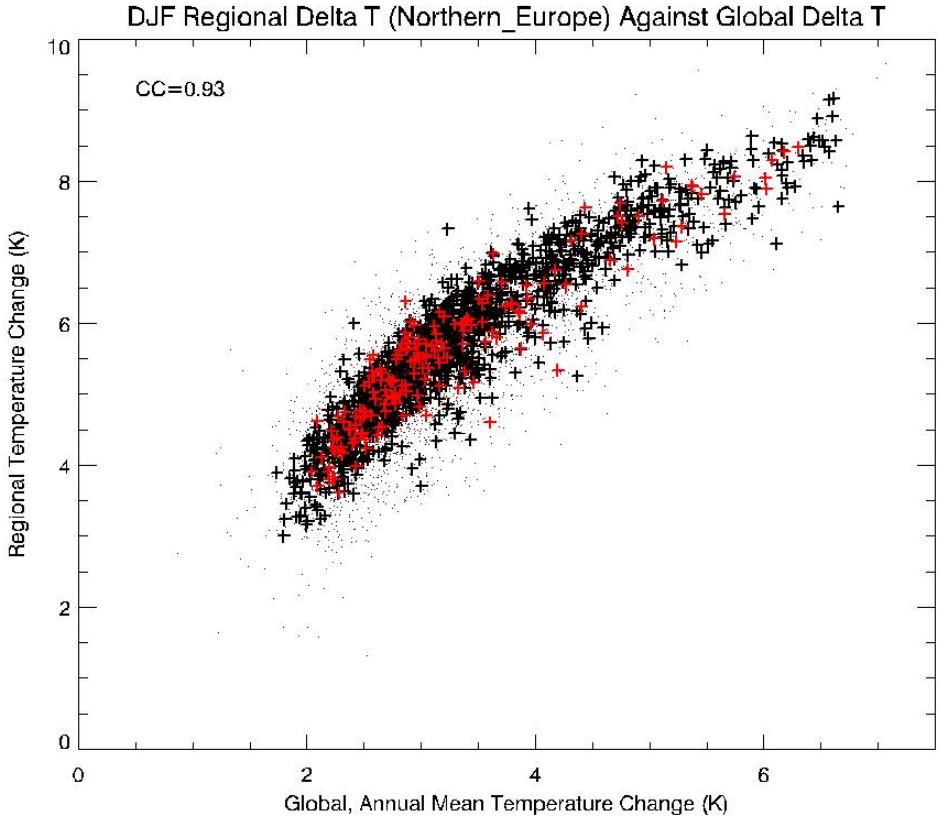
Min ICE	Total points
1	6203
4	1594
5	996
6	563
7	259
8	91

3 - At least four member Initial condition ensemble members



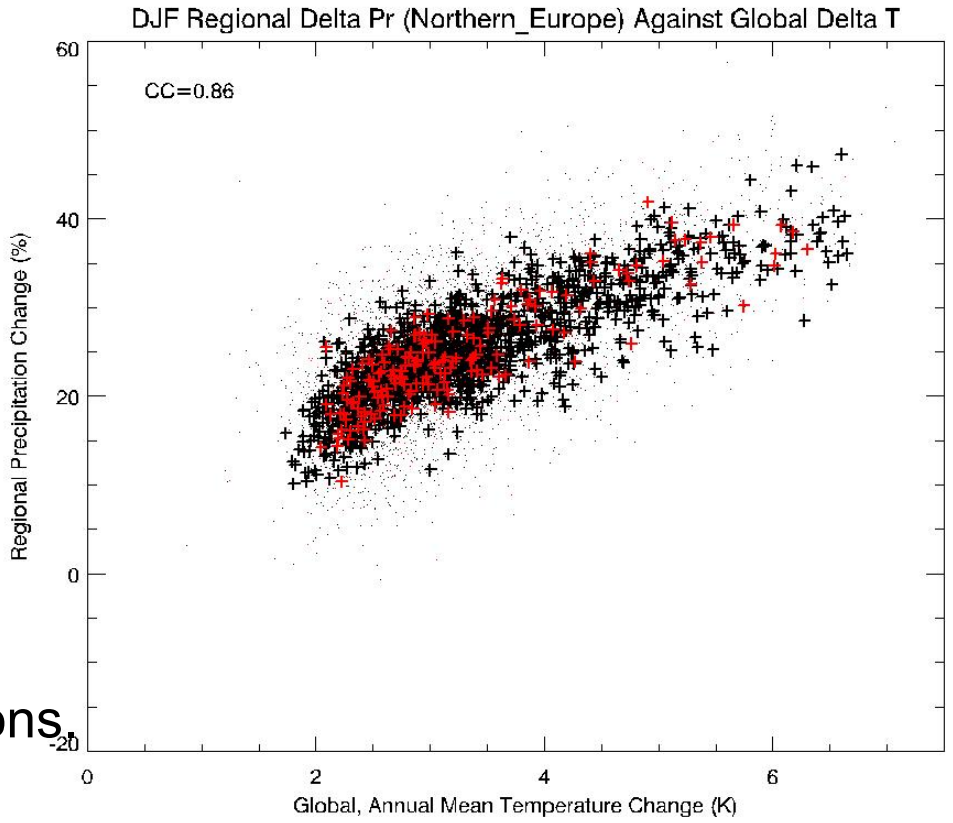


4 – Culling by Atmosphere/Ocean Heat Flux



Challenge 3: Model culling

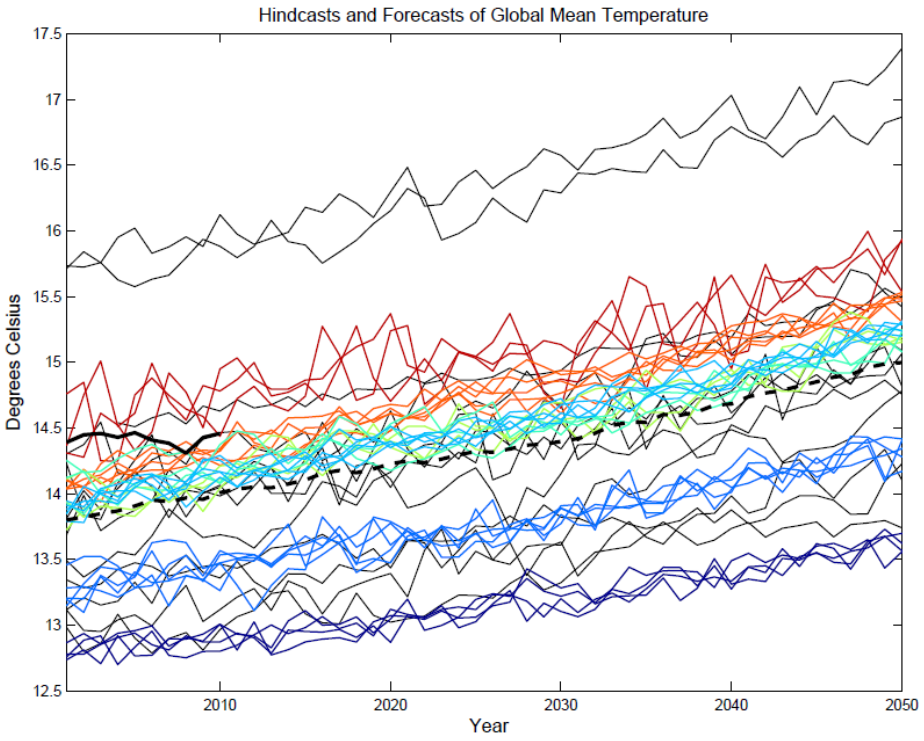
- How do we decide which models are so bad they should not be studied?



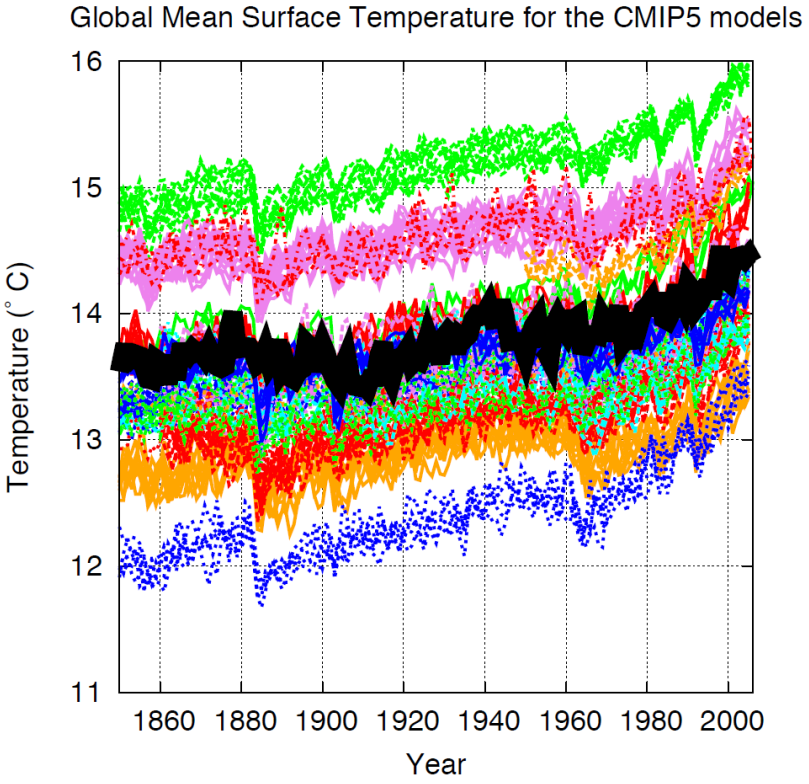
Remember:

- This is a complex non-linear system.
- All models are inconsistent with observations.
- So what is “just too bad”?

Evaluating Model Quality / Model Weighting

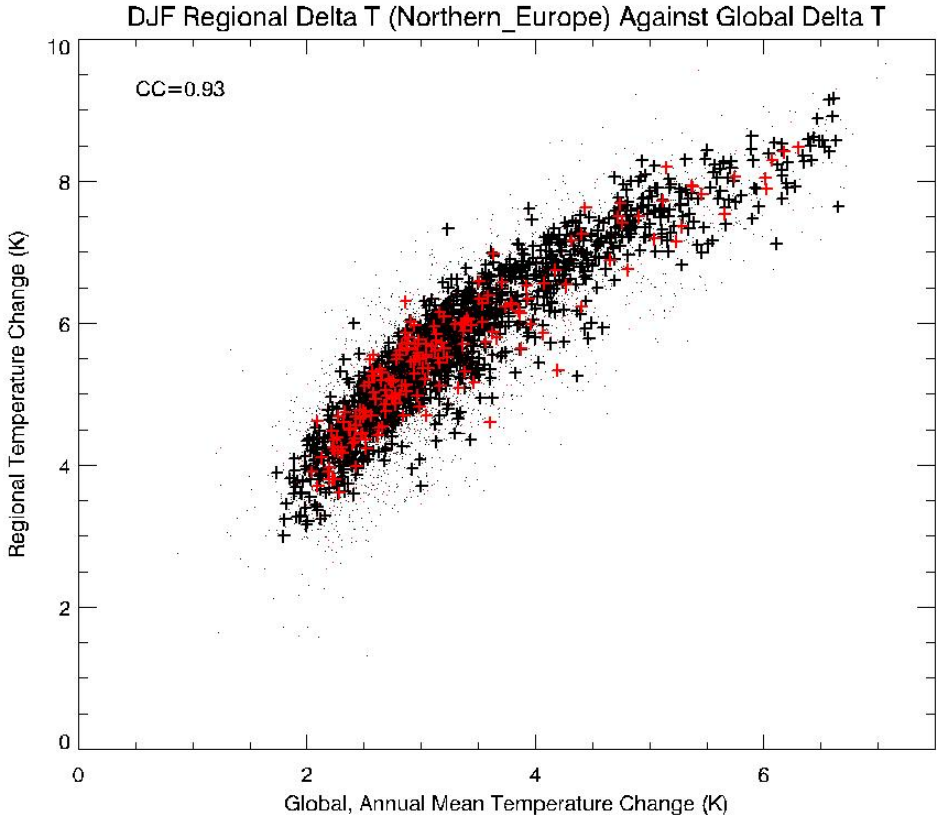


Acknowledgement: Ana Lopez



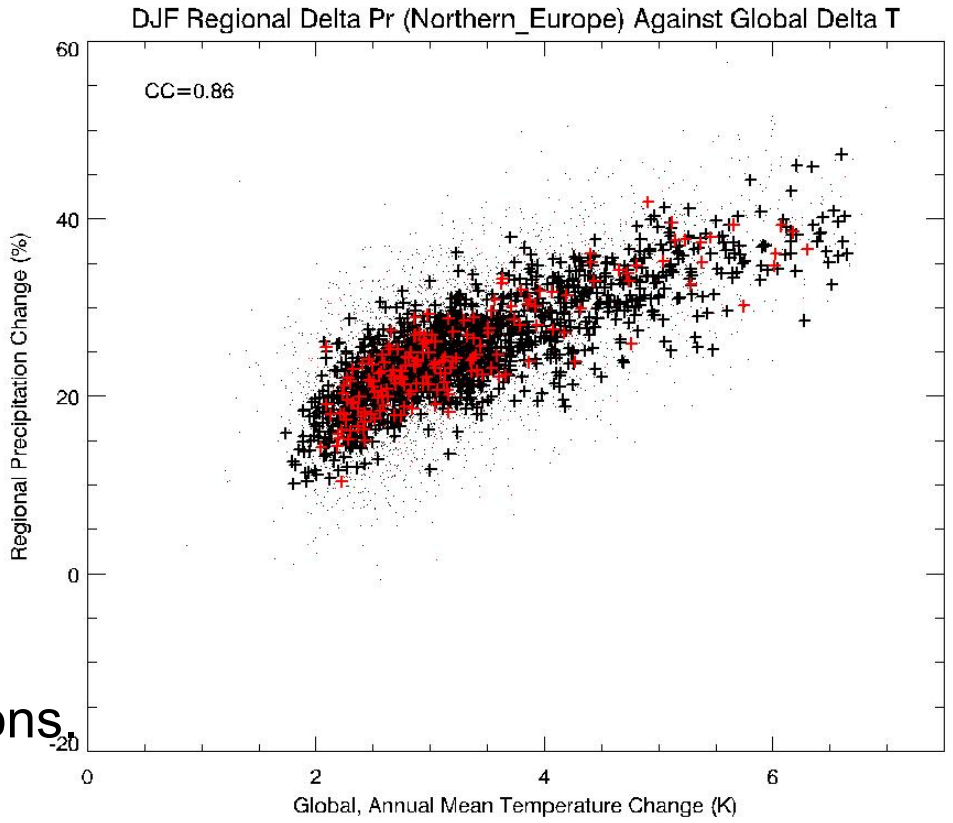
Acknowledgement: Emma Suckling

4 – Culling by Atmosphere/Ocean Heat Flux



Challenge 3: Model culling

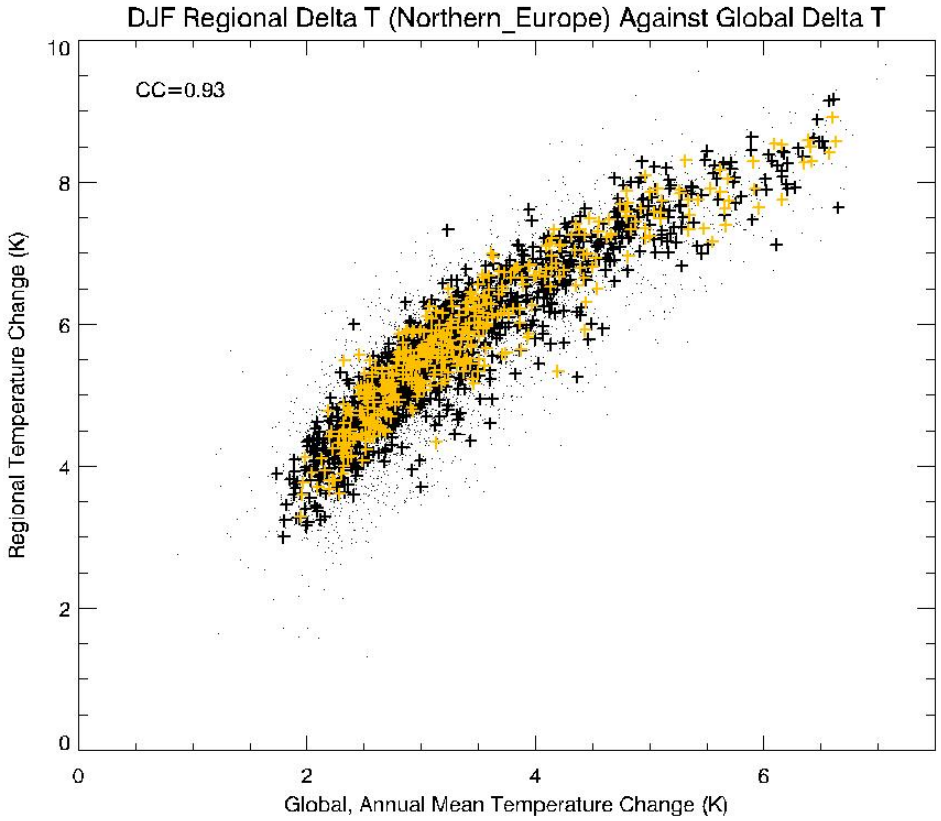
- How do we decide which models are so bad they should not be studied?

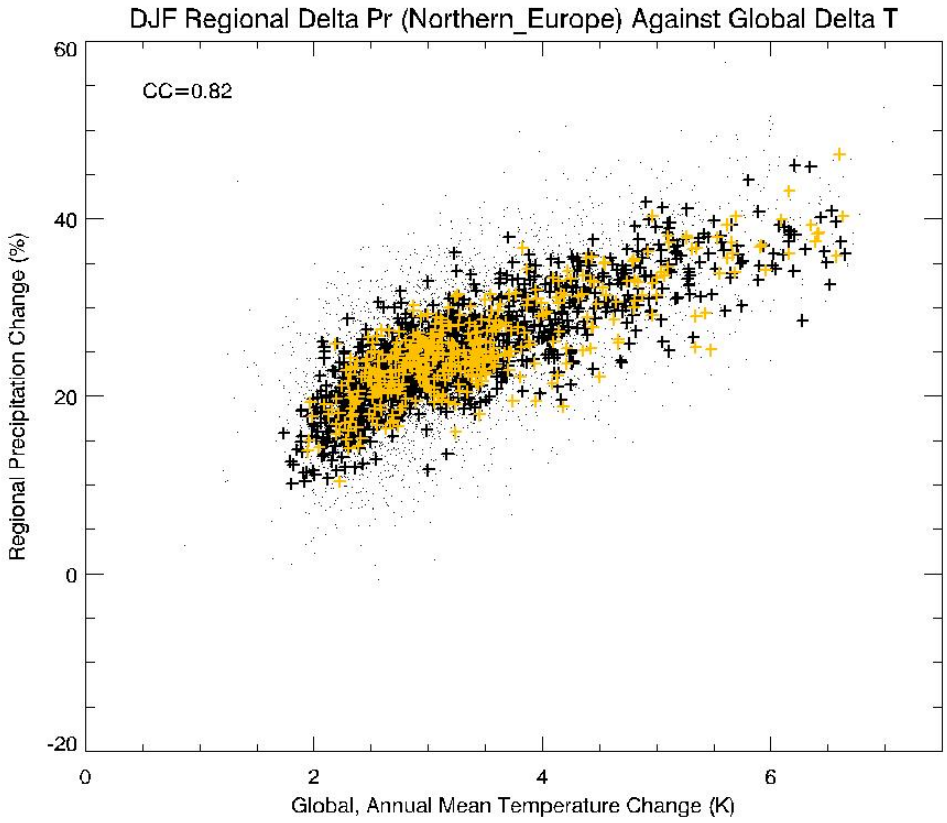


Remember:

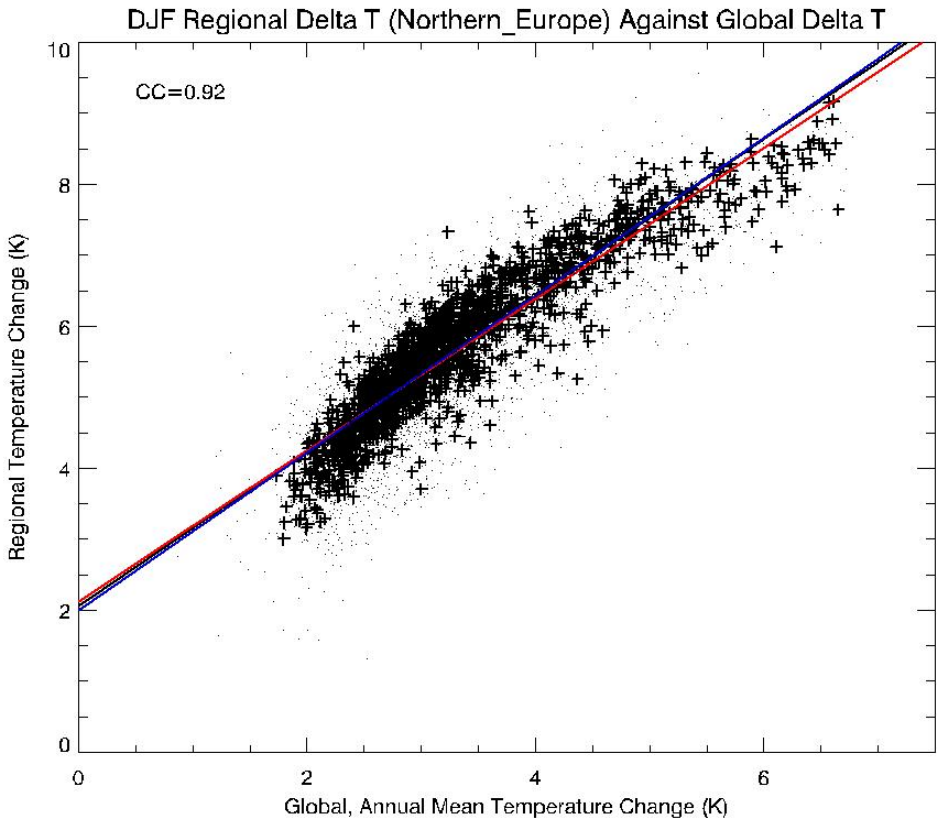
- This is a complex non-linear system.
- All models are inconsistent with observations.
- So what is “just too bad”?

6 – Culling by entrainment coefficient



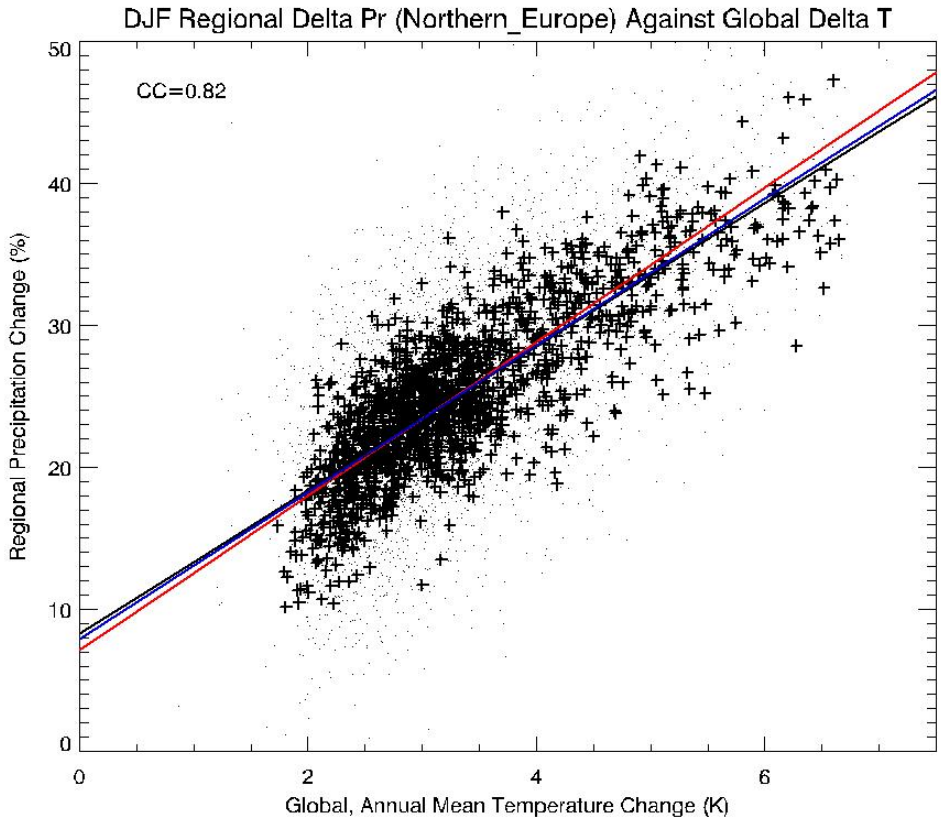


7 – Linear Fits

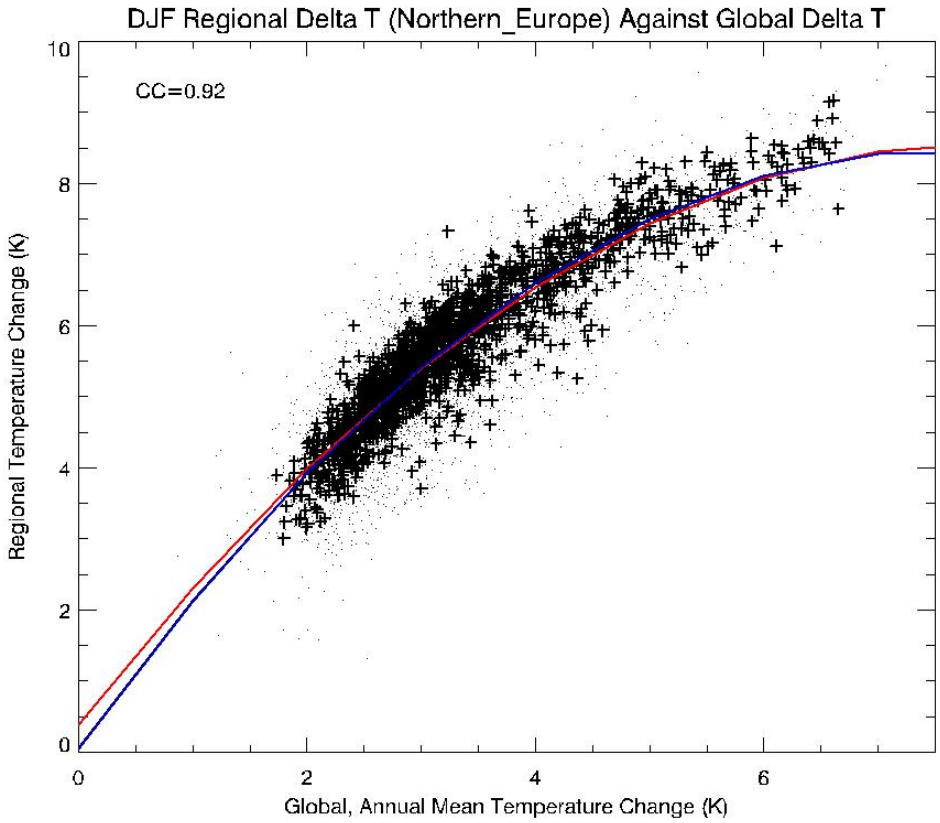


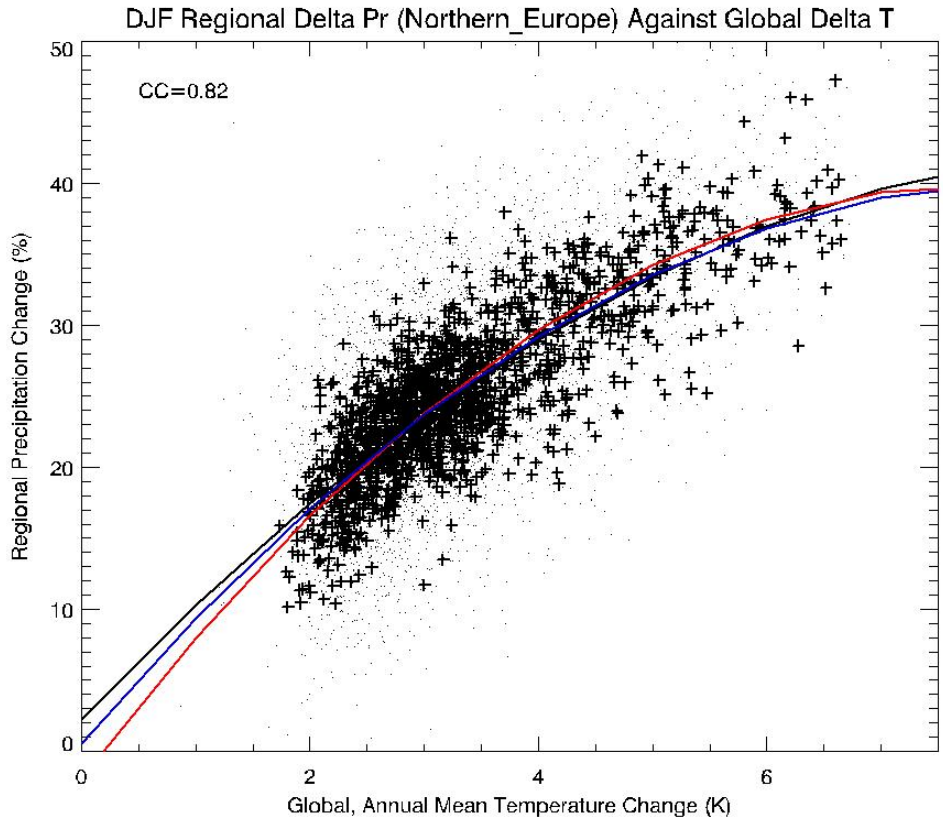
Challenge 4: What should we take from a fit across different models mean?

- They are neither different states of the same model nor independent models.

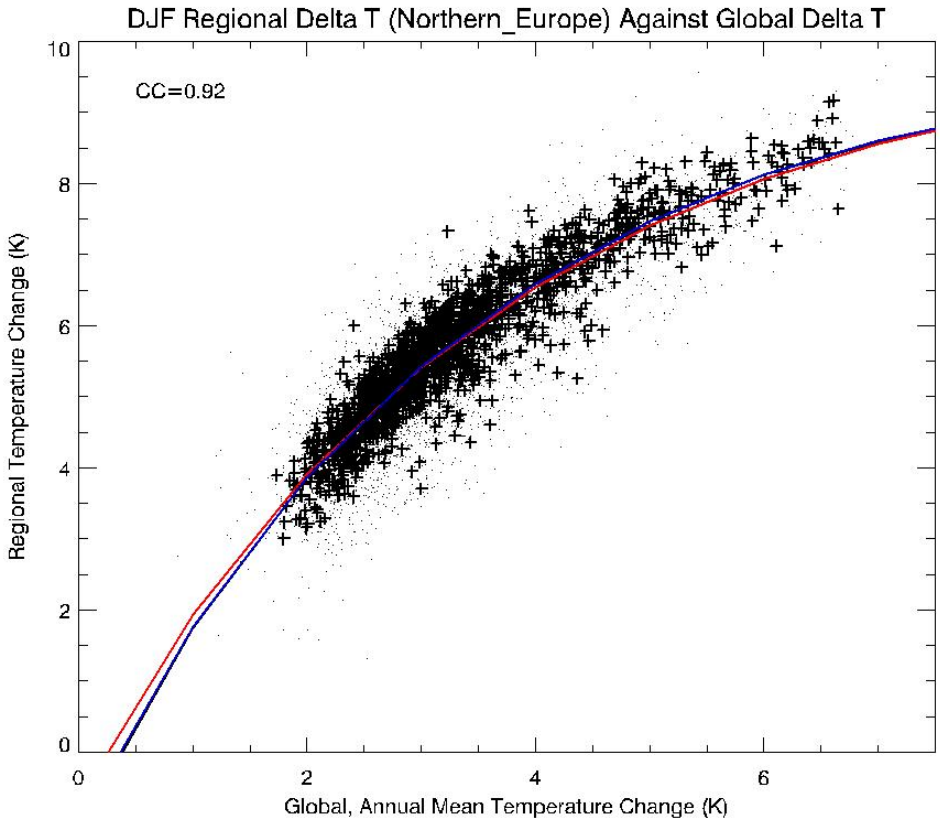


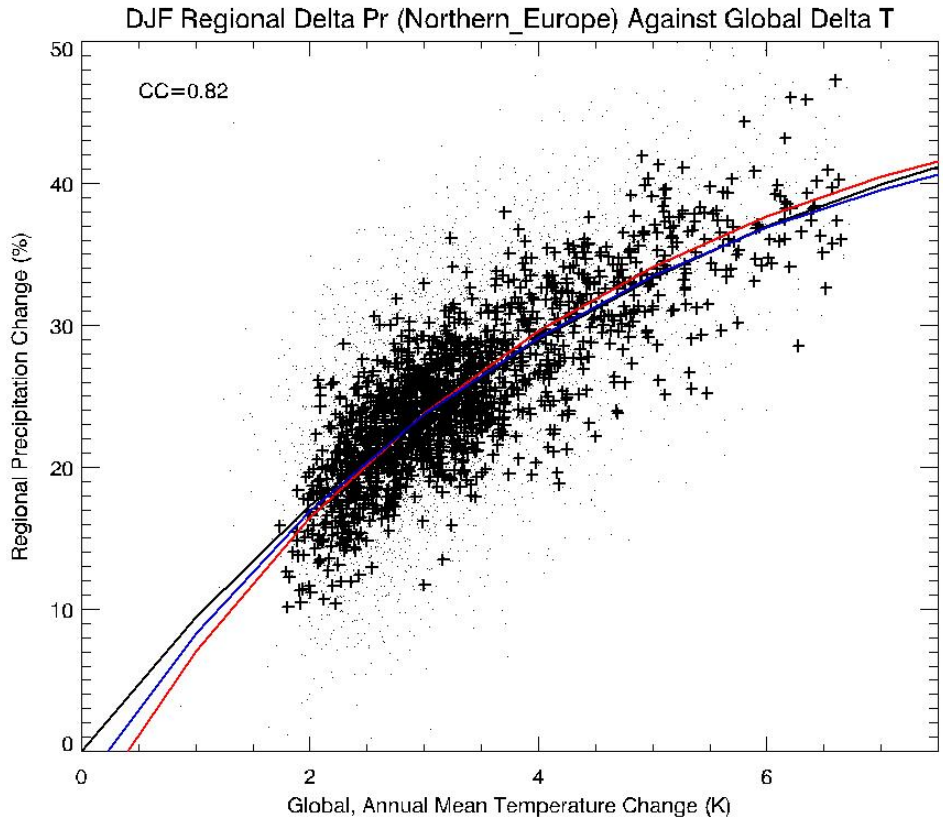
12b – Polynomial Fit



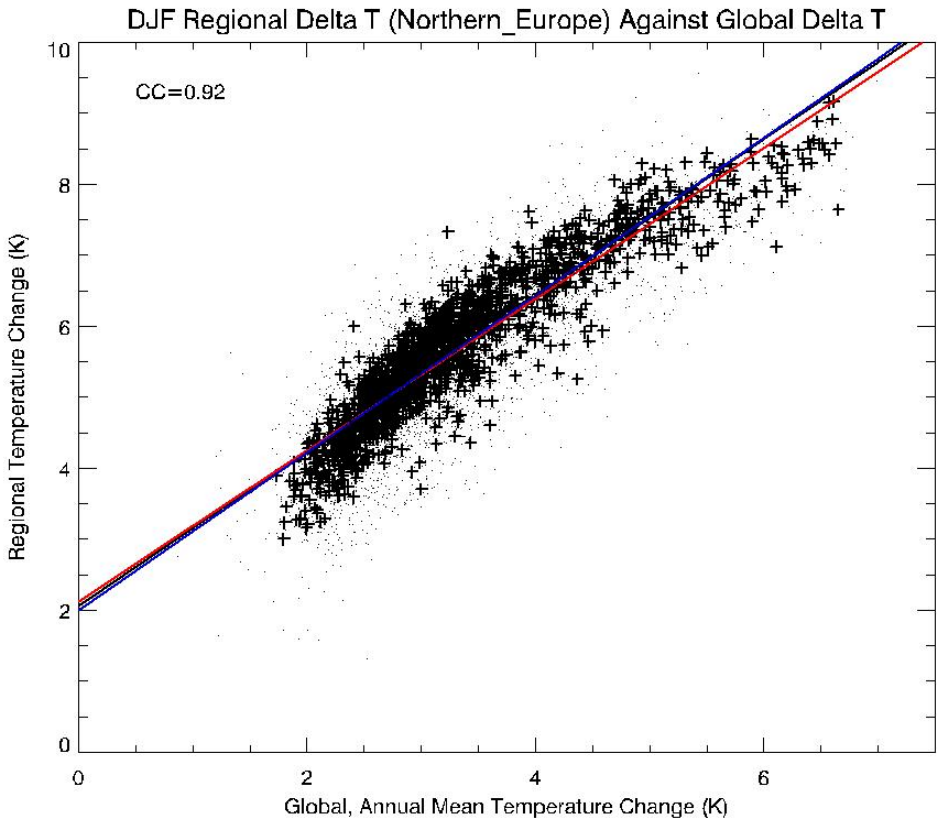


8b – Exponential Fit



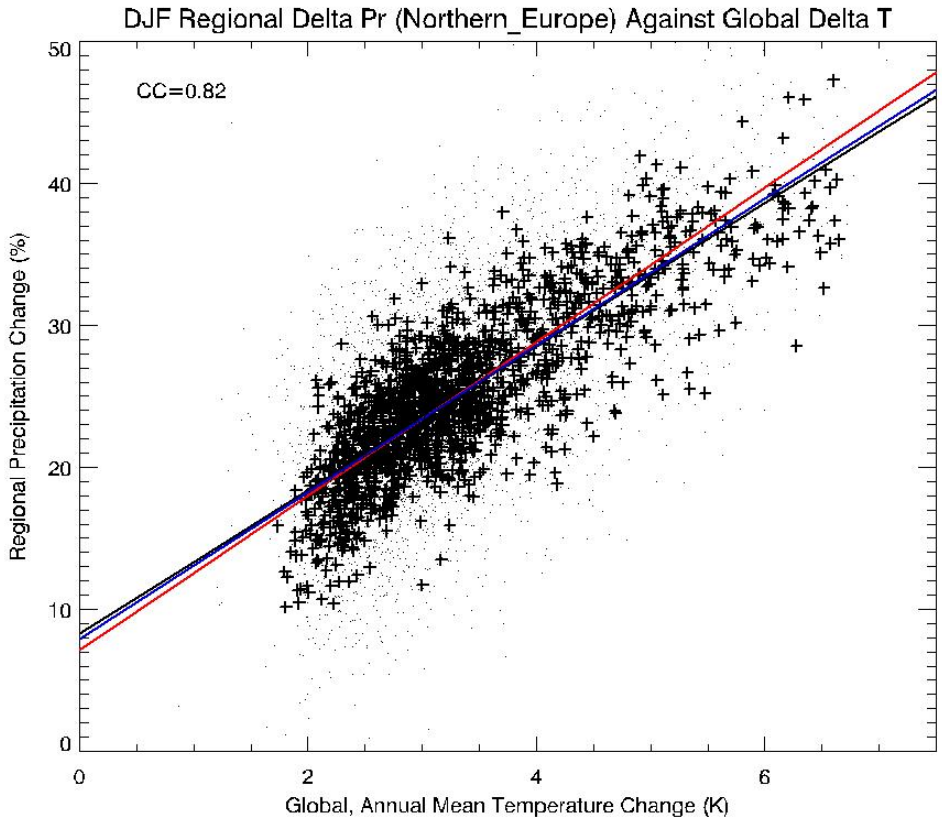


7 - Are They Good Fits?



Challenge 1: Coping with lack of independence.

Challenge 5: Evaluating model dependence.
(On inputs rather than outputs?)



χ^2 probability assuming all models independent:

100.00%(temperature), 100.00%(precip)

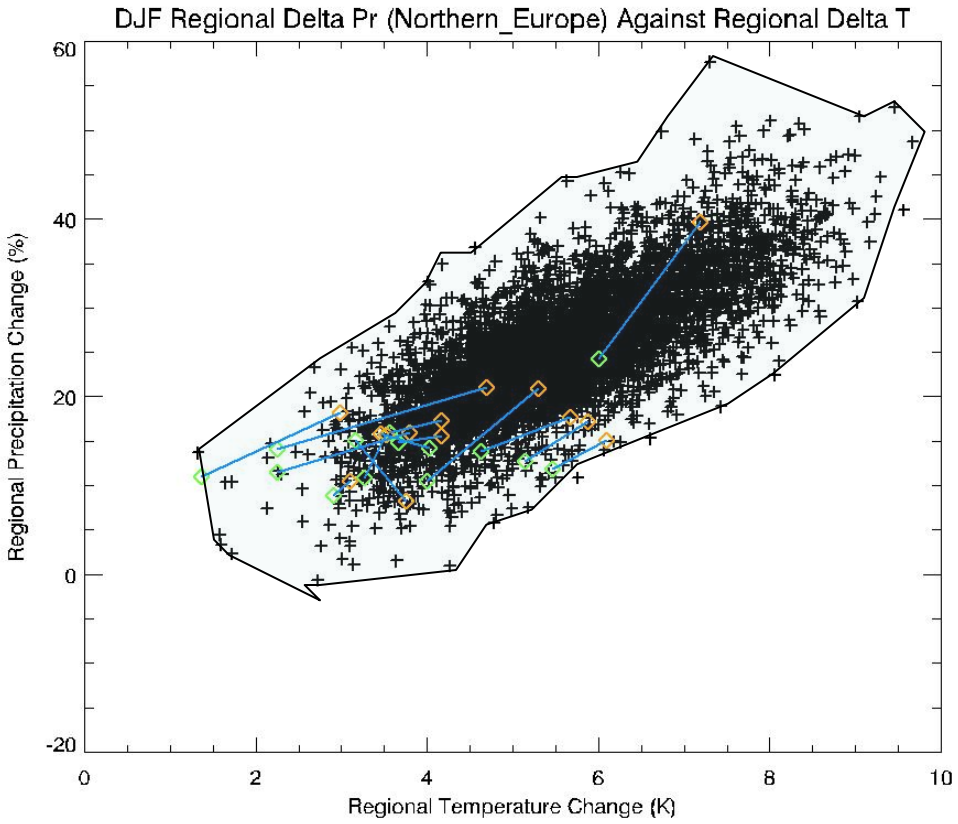
χ^2 probability assuming no. of independent models is ¼ of total:

0.000% (temperature), 0.001%(precip)

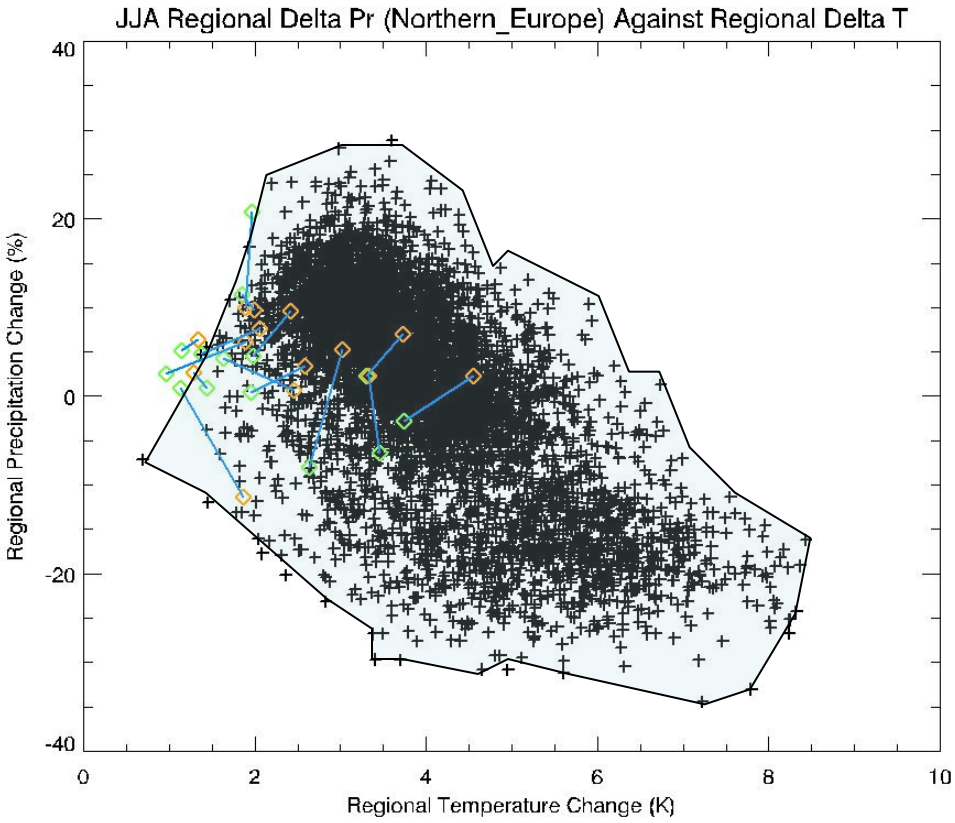
Challenges in Interpretation

- Independence:
  Model versions are not independent samples of possible models. So how do we statistically interpret them?
  - Model diversity .vs. real world probability.
    - There is no reason to expect the density of points to reflect confidence, likelihood or probability in the real world.
  - The shape of model space is arbitrary.
- Model culling or weighting.
  - How do we decide which models are so bad they should not be studied?
  - Remember - this is a complex non-linear system. In terms of predicting the future, under changes in forcing, there is no value in selecting models which simulate our region/variable of interest well if it gets other regions/variables badly wrong.
  - All models are inconsistent with observations.
  - So what is “just too bad”?
  - What data there is is for a different state of the system and has already been used i.e. it is in-sample.
- In-sample ensemble analysis.
- Expert Opinion:
  - Most climate scientists are climate modellers. The models are themselves the number one source of information for “experts”. Isn’t this all too self-referential?
  - Do experts have probabilistic understanding.
- Extrapolation.

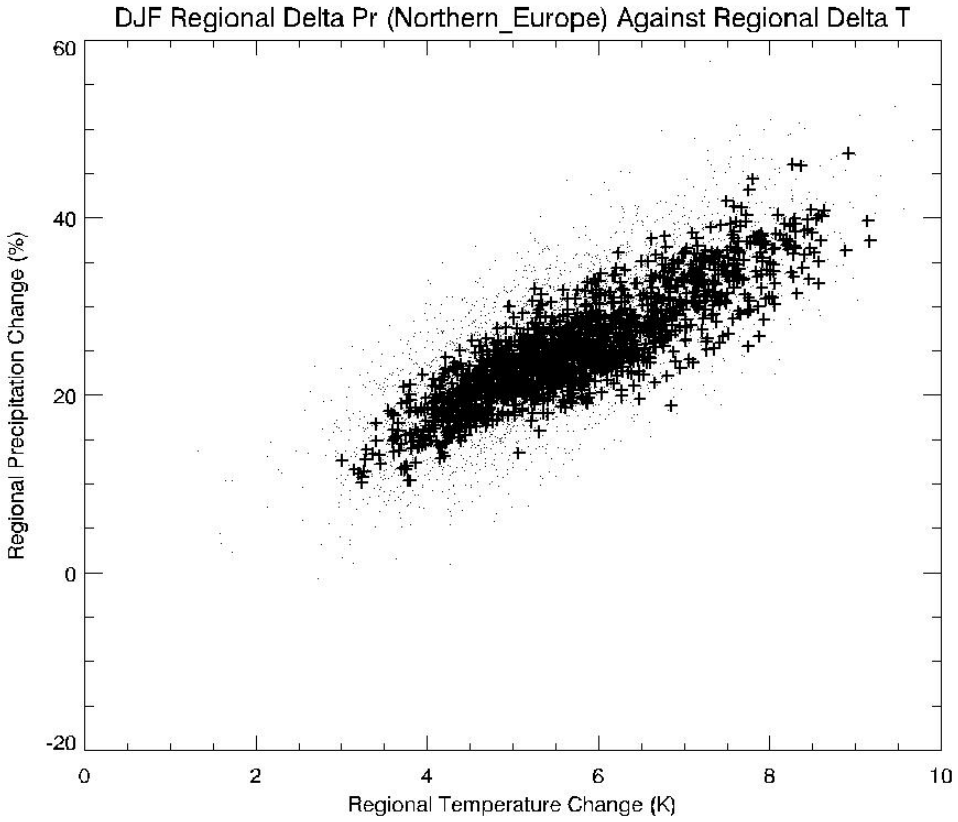
Without independence;
with or without a credible weighting or culling strategy;
the most we have is a domain of possibility, a non-discountable envelope

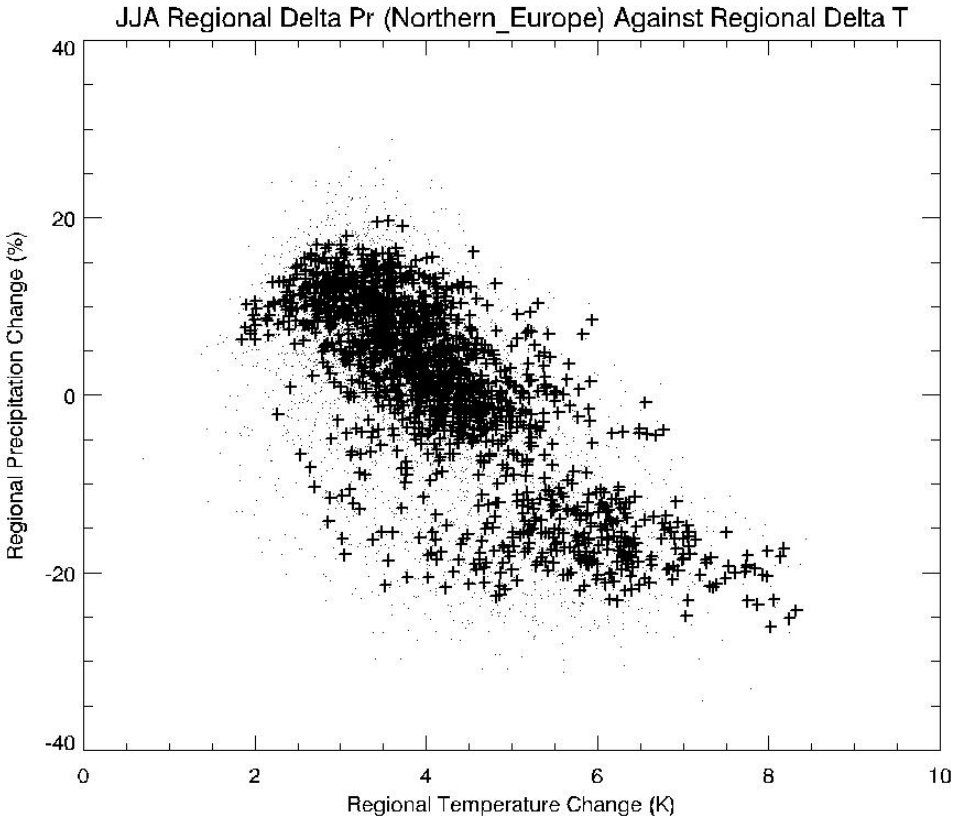


- What does this say about what's inside and what's outside the domain?

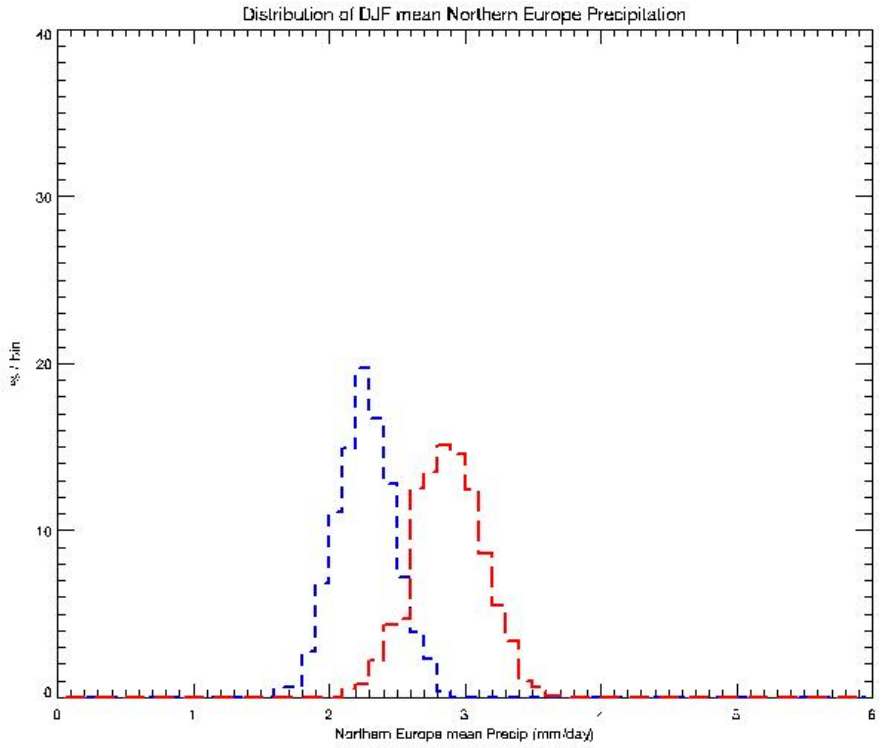


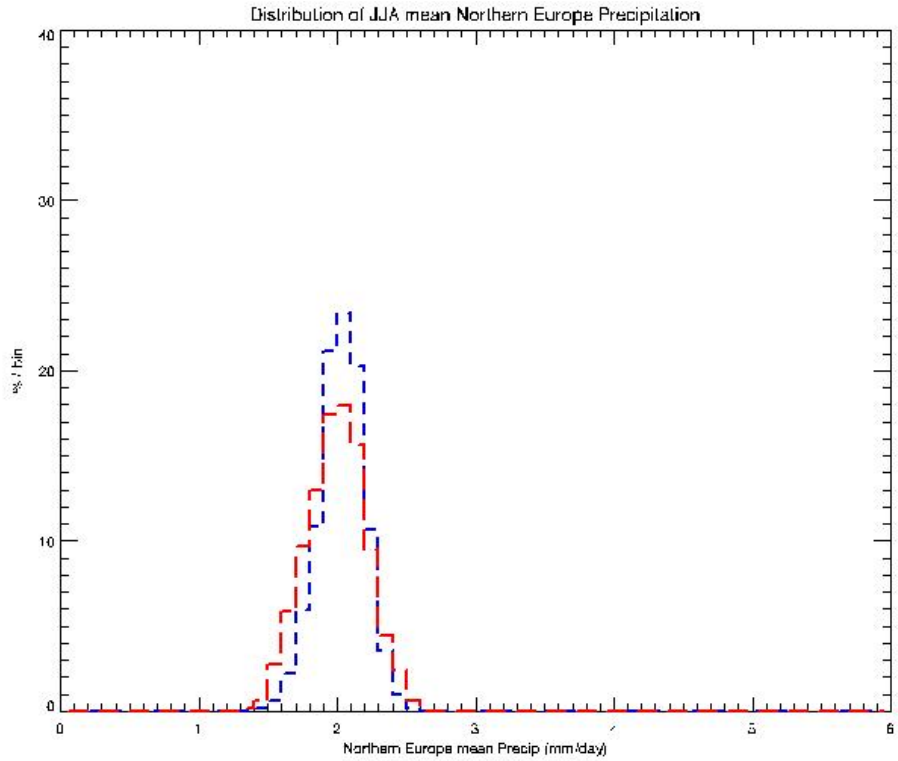
Separating the Consequences of Model Uncertainty and Initial Condition Uncertainty Shrinks the Domain But for any Practical Decisions Initial Condition Uncertainty Must Then Be Added On.





And Initial Condition Uncertainty is Not Small
[NB These are distributions of seasonal values for individual years rather than 8 year means]





Some of the Challenges

- Lack of independence.
- In-sample ensemble analysis.
- How do we cull ensembles? How can weighting make sense when all our models are so bad?
- What do relationships across different model means mean?
- How do we evaluate model independence – on inputs rather than outputs?

- Discussion points:
  - Is our aim to **reduce** uncertainty?
  - Do experts have probabilistic information/? (see Milner et al. 2013)

How Should We Design Ensembles?

- Design ensembles to push out the model domains
  - Climate models are a tool for better understanding. Diversity:
    - supports better differentiation of the plausible from the implausble and
    - Encourages broad questioning of results so we don't fool ourselves (and society) into thinking certain behaviour is unlikely/impossible just because our models don't show it.
  - If used as quantitative predictors at all then we begin to get constraints from what they can **not** do.
- Ensembles which substantially explore uncertainty in the transient response.
  - Design of such ensembles is a priority.
  - Computer capacity
  - A multi-disciplinary debate addressing the value of ensemble size c.f. model resolution.
- Understanding how such ensembles provide value for:
  - Scientific understanding.
  - Model development.
  - Climate predictions.

References

- Stainforth, D. A., Allen, M. R., Tredger, E. R., and Smith, L. A., **Confidence, uncertainty and decision-support relevance in climate predictions**. Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences 365 (1857), 2145 (2007).
- Stainforth, D. A, T.E. Downing, R. Washington, A. Lopez, M. New. **Issues in the interpretation of climate model ensembles to inform decisions.** Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences 365 (1857), 2163 (2007).
- Daron, J. D. & Stainforth, D. A. On predicting climate under climate change. *Environ. Res. Lett.* **8** (2013)
- Frigg, R., S. Bradley, H. Du, & L. A. Smith, *Laplace's Demon and the Adventures of His Apprentice*, Philos. Sci. (2014).
- Smith, L. A., **What might we learn from climate forecasts?** Proceedings of the National Academy of Sciences of the United States of America 99, 2487 (2002).