A Statistical Perspective on Uncertainty Quantification

Murali Haran

Department of Statistics, Pennsylvania State University

Network for Sustainable Climate Risk Management (SCRiM) Summer School, 2019.

What We Will Discuss

- Why uncertainty quantification is central to climate risk management and, more broadly, to science
- Examples that focus on the importance of uncertainty quantification when studying climate models, and particularly when climate projections are made using models
- A few basic ideas related to statistical inference

Uncertainty Quantification and Information

- Statistical inference provides a rigorous way to translate observations into information.
- Example: Look at time series of temperatures to obtain trends. Are these trends signal or noise (significant or insignificant)? Understanding this involves the variability of the system, observations, limits of our knowledge.
- Another example: when interpolating across space, across parameters or scenarios (when studying models), use statistics to quantify interpolation uncertainty/error.
- What climate models can/cannot say about true system.
- Closely related: UQ provides a rigorous way to ascertain *how much* information the observations provide about the system of interest. Can quantify “the value of information”.

Quantifying Uncertainty

- Uncertainty is not the same as not knowing.
- A lack of certainty is not a justification for inaction. The notion of risk is formulated precisely to handle decision making under uncertainty.
- Describing uncertainties carefully is central to the scientific enterprise. For instance, “Without uncertainty quantification, it is easy to dismiss climate (computer) models.” – A. O’Hagan.

Climate Change and Risk

Polder dyke, Netherlands (from John Elk III, lonelyplanet.com)

Risk

- We need probability/statistics to define and quantify risk.
- Risk associated with an action = expected cost (or “loss”) for that action, “expected” = weighted average
  - Sum over { probability of outcome $\times$ cost of that outcome}
- For a particular policy, example of outcomes:
  - strength of the Atlantic Meridional overturning circulation or “AMOC”: weakening? stable?
  - global sea level rise: 2 metres? more? less ?
- To study economic impacts: relate outcome to impact. Example: sea level rise of x metres will cost \$$y$.
- Common to misunderstand and confuse probability and risk. E.g. low probability-high impact events are thought of as being more "likely" than low probability-low impact events.

Learning about Risk

- What is probability of sea level rise of 2m. in 2100 if:
  (A) Carbon emissions grow at same rate (“business as usual”)
  (B) Carbon emissions are controlled by a policy
- Of particular interest: low probability-high impact events. For example sea level rise of 2 metres may be a relatively low probability event but extremely expensive. "Tails" of probability distributions are important.

How do we learn about the probability of each outcome while accounting for uncertainties?

Learning about Risk through Statistical Methods

Risk assessment based on climate projections involves:

(1) Combining information from climate models and observations.

(2) Uncertainty quantification: for honest assessment of risk, critical to incorporate information about how certain or uncertain we are about various aspects of the climate projections.

(3) Addressing technical challenges related to the size of the data sets involved.

Novel statistical methods have been/are being developed to address the above issues.

Types of uncertainty

(1) Aleatoric: stochasticity (randomness) in the universe. Example: if we knew a coin was fair, still would not know if a particular toss would yield heads or tails.

(2) Epistemic: uncertainty regarding our knowledge. Example: the weight of a particular coin (nickel/5c) is fixed but our knowledge about the weight is uncertain. If we knew the weights of 20 other coins (nickel/5c), we could make a better guess (reduced uncertainty).

Statistical models may be used to account for both.

Categories of Uncertainties for Climate Science

An approach relevant to climate policy is as follows (cf. Smith and Stern, 2011). These categories are not mutually exclusive.

- Imprecision (Knightian risk): related to outcomes which we do not know precisely, but for which we believe robust, decision-relevant probability statements can be provided.
- Ambiguity (Knightian uncertainty): related to outcomes (known, unknown or disputed), for which it is difficult to make probability statements. Sometimes “scenario uncertainty” or uncertainty in an estimated probability (“second-order uncertainty”).

Categories of Uncertainties (continued)

- Intractability: related to inability to formulate/execute relevant computations.
- Indeterminacy: related to quantities relevant to policy-making for which no precise value exists. E.g. a model parameter that does not correspond to an actual physical quantity, honest differences in point of view due to differences in objectives/values

Climate Model Uncertainties

Models: systems of equations, simulated on a computer. Used to understand interplay between processes that drive system, e.g. ocean circulation, ice sheet models

1. Models can never fully describe climate system. Structural uncertainty: Which model features to prioritize?
2. Internal/natural climate variability: interactions between components of the climate system can result in variability in the climate system that is "internal" to the system.
3. Boundary or initial condition uncertainty.
4. "Forcings" uncertainty, e.g. uncertainty about emissions.
5. **Parameter uncertainty**: parameters ("dials" in the computer model) may be uncertain.
6. Observations: measurement, interpolation uncertainty

Sources of Information on Climate

Lots of great climate information, lots of complications.

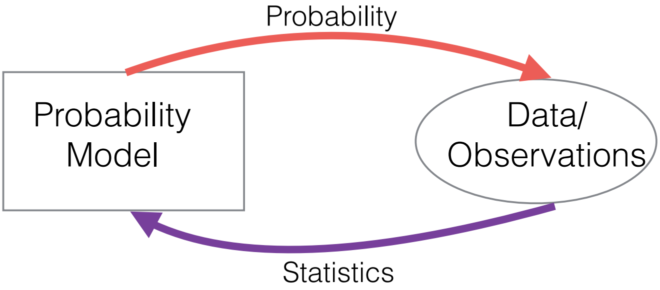
Studying the past, present, future of climate involves using various kinds of data/sources of information. Rough taxonomy

- "Direct" measurements
  - Temperature data with thermometers, tide gauge data, sea surface temperature data from buoys, satellite data, radar data
- Indirect measurements
  - ("Very indirect") Temperature proxies, e.g. ice cores, tree core; ice sheet thickness
- Output from physical models
  - General (global-scale) circulation models, regional climate models
  - Smaller-scale, e.g. ice sheet models

Complications with Climate Information

- Most information in climate science data does not neatly fall into one of the above categories
- Climate information lies on a continuum. Invariably some processing/interpolation/modeling is involved
  - Radar data, satellite data
  - "Reanalysis" data; information obtained from data assimilation (combining observations with model output)
- The centrality of statistical modeling in climate science: allows us to rigorously quantify the various sources of uncertainty in the information we use
- When drawing scientific conclusions that use these data, statistical modeling/thinking allows us to carefully account for how the information/data were obtained

Probability and Statistics



Probability Model = Deterministic Model + Randomness/Errors

The Language of Probability

Formulating a model that describes what we observe (data), *while accounting for various uncertainties*

- Probability distribution (special cases: probability density function, probability mass function), call it $f(x;\theta)$
- Examples when data are in the form of scalars:
  - Normal distribution ("bell curve", Gaussian distribution). Later: need to find parameters (mean μ, variance σ^2) to "fit" a particular data set
  - Exponential distribution. Later: need to find parameter (λ) to "fit" a particular data set

Advanced Probability Models

In climate science, common to have to model complicated phenomenon, account for relationships among the observations

- Time series models: data have time associated with them, e.g. daily temperature data. Often related to each other (tomorrow's temperature looks more like today's temperature than the temperature 4 weeks ago); also exhibit seasonal cycles etc.
- Spatial models: data have locations associated with them, e.g. precipitation data across Pennsylvania. May have to model interesting relationships in space.

Uncertainty and Hierarchical Models

- We (hopefully) agree that there are multiple kinds of uncertainty we need to incorporate into our analysis
- How do we do this in practice?
- Elegant framework: hierarchical modeling
- Write down a probability model for anything that is uncertain
- A Bayesian approach makes this easy. Eg of books:
  - Hobbs & Hooten (2015) Bayesian Models: A Statistical Primer for Ecologists
  - Sivia and Skilling (2008) Data Analysis: A Bayesian Tutorial

A Simple Hierarchical Model

- For some species, let Y = number of eggs that survive/lead to viable offspring
- Build a hierarchical model for Y in stages:
  - Let X be the number of eggs laid by an individual

A Simple Hierarchical Model

- For some species, let Y = number of eggs that survive/lead to viable offspring
- Build a hierarchical model for Y in stages:
  - Let X be the number of eggs laid by an individual
  - Let Y be the number of these eggs (out of X) that survive

A Simple Hierarchical Model

- For some species, let Y = number of eggs that survive/lead to viable offspring
- Build a hierarchical model for Y in stages:
  - Let X be the number of eggs laid by an individual
  - Let Y be the number of these eggs (out of X) that survive
  - Assume each egg has equal probability of survival, p, and each egg's survival is independent .

A Simple Hierarchical Model

- For some species, let Y = number of eggs that survive/lead to viable offspring
- Build a hierarchical model for Y in stages:
  - Let X be the number of eggs laid by an individual
  - Let Y be the number of these eggs (out of X) that survive
  - Assume each egg has equal probability of survival, p, and each egg's survival is independent .
  - Probability model for X: Poisson(λ)

A Simple Hierarchical Model

- For some species, let Y = number of eggs that survive/lead to viable offspring
- Build a hierarchical model for Y in stages:
  - Let X be the number of eggs laid by an individual
  - Let Y be the number of these eggs (out of X) that survive
  - Assume each egg has equal probability of survival, p, and each egg's survival is independent .
  - Probability model for X: Poisson(λ)
  - Probability model for Y *given* X, $Y|X$ is Binomial(X, p)
- Two random variables, X, Y, two parameters λ, p
- Use statistical methods to find λ, p for a data set
- Possible that you only observe Y, in which case X is "latent" or unobserved

General Hierarchical Framework

Common to use hierarchical framework of Mark Berliner (1994)

- **Prior** (θ) model ($p(\theta)$): describes assumptions about parameters of the model. Uncertainties described via distributions

General Hierarchical Framework

Common to use hierarchical framework of Mark Berliner (1994)

- **Prior** (θ) model ($p(\theta)$): describes assumptions about parameters of the model. Uncertainties described via distributions (epistemic uncertainty)
- **Process** (X) model ($g(X|\theta)$): describes the model for the process of interest, example a dynamical system that you probably cannot observe directly, example the evolution of an ice sheet

General Hierarchical Framework

Common to use hierarchical framework of Mark Berliner (1994)

- **Prior** (θ) model ($p(\theta)$): describes assumptions about parameters of the model. Uncertainties described via distributions (epistemic uncertainty)
- **Process** (X) model ($g(X|\theta)$): describes the model for the process of interest, example a dynamical system that you probably cannot observe directly, example the evolution of an ice sheet
- **Data** (Y) model ($f(Y|X,\theta)$): probability model that describes the observation process, e.g. measurement error and other complications

General Hierarchical Framework

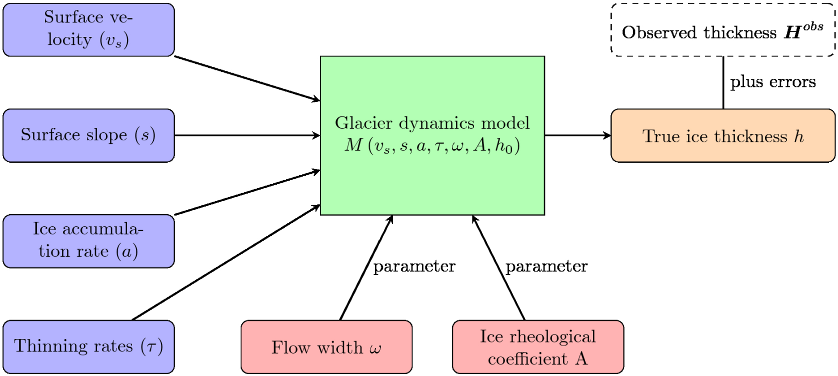
Common to use hierarchical framework of Mark Berliner (1994)

- **Prior** (θ) model ($p(\theta)$): describes assumptions about parameters of the model. Uncertainties described via distributions (epistemic uncertainty)
- **Process** (X) model ($g(X|\theta)$): describes the model for the process of interest, example a dynamical system that you probably cannot observe directly, example the evolution of an ice sheet
- **Data** (Y) model ($f(Y|X,\theta)$): probability model that describes the observation process, e.g. measurement error and other complications

A systematic way to view scientific modeling. Model for data Y is based on models for: (1) $Y|X$, (2) $X|\theta$, (3) θ

Hierarchical Models

An approach to build complex models/relationships among data



The Language of Statistics

Once we have a probability model for our data, rigorous ways to learn about the parameters of the model:

- Likelihood function (maximum likelihood inference)

  Likelihood = probability distribution function with the data values fixed at observations

  Maximize this likelihood function with respect to parameters
- Posterior distribution (Bayesian inference)

  Posterior = *conditional* probability distribution

  look at the distribution of our parameters *given* the data

Statistical Methods: Modeling

An approach to formalize the statistical methodology:

(1) Modeling of the system:

- If possible, establish a forward model that describes the dynamics (physics) of the system of interest. This is a *mathematical model*, which may be implemented through computer code.
- Add potential errors, biases to the above to account for various uncertainties. This now describes a *statistical* model for the system.
- Denote observations of this system by **Z**, parameters of the forward model by $\boldsymbol{\theta}$, statistical parameters by ξ.
- A "realization" (a particular instance) of **Z** is then obtained from this probability distribution, $f(\mathbf{Z}; \boldsymbol{\theta}, \boldsymbol{\xi})$

(Modulo computational challenges) it should now be possible to simulate **Z** using various values of $\boldsymbol{\theta}, \boldsymbol{\xi}$. *Built-in variability.*

Statistical Approach for UQ: Inference

(2) Likelihood-based inference: General approach to take an observation of **Z** and model above and (i) estimate $\boldsymbol{\theta}, \xi$, (ii) approximate uncertainty about $\boldsymbol{\theta}, \xi$.

- Likelihood function, $\mathcal{L}(\boldsymbol{\theta}, \xi) \propto f(\mathbf{Z}; \boldsymbol{\theta}, \xi)$ with **Z** fixed at observed values. $\mathcal{L}(\boldsymbol{\theta}, \xi)$ function of *only* parameters.
- Maximize $\mathcal{L}(\boldsymbol{\theta}, \xi)$ w.r.t. $\boldsymbol{\theta}, \xi$ to obtain maximum likelihood estimates, $\hat{\boldsymbol{\theta}}, \hat{\xi}$
- Curvature of $\mathcal{L}(\boldsymbol{\theta}, \xi)$ at $\hat{\boldsymbol{\theta}}, \hat{\xi}$ approximates uncertainties. (Argument is asymptotic, i.e., as sample size tends to infinity.)
- Alternative approach: Bayesian inference.

Statistical Approach for UQ: Inference

(2*) Bayesian inference:

- Begin by specifying **prior** distribution for $\boldsymbol{\theta}, \xi$: What we thought they were before observing **Z**, described as a distribution, $p(\boldsymbol{\theta}, \xi)$.
- What we know about $(\boldsymbol{\theta}, \xi)$ is described by posterior distribution (conditional on data) $\pi(\boldsymbol{\theta}, \xi \mid \mathbf{Z}) \propto \mathcal{L}(\boldsymbol{\theta}, \xi) p(\boldsymbol{\theta}, \xi)$.
- Instead of optimization, computation is usually Monte Carlo integration using Markov chain Monte Carlo (MCMC):
  - Simulate from the distribution $\pi(\boldsymbol{\theta}, \xi \mid \mathbf{Z})$.
  - Based on these simulations, approximate properties of this distribution.

Model Building and Evaluation

Reminder:

- Important: Steps 1 (Model Specification) and Step 2 (Statistical Inference) are not simply done once.
- After statistical inference/model fitting (Step 2), need to return to Step 1 to evaluate whether fitted model, error structures etc. are appropriate.
- Step 1 is difficult and important, uses both scientific and statistical expertise. May have to iterate many times before reasonable model and inference obtained.

Uncertainty Quantification and Inverse Problems

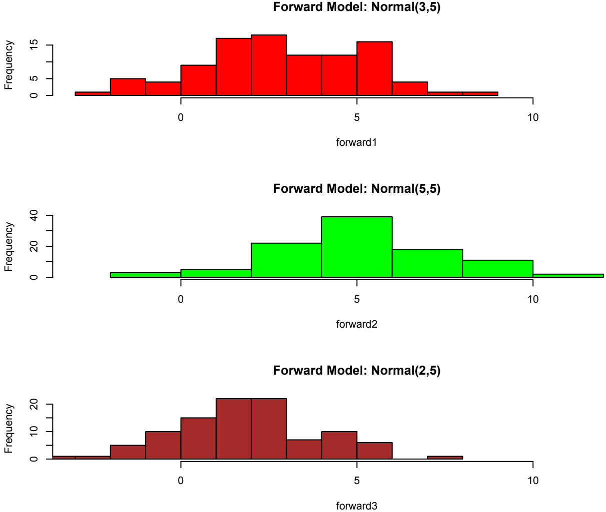
- One could think of statistics as being about "solving" an inverse problem. Invariably these inverse problems (especially in climate science!) do not result in unique solutions. Noisy measurements and inaccuracies in data gathering contribute to this partially but even with perfect data, it is rare to have unique solutions.
- Many different solutions (solutions may be models, parameters) to the same problem. Uncertainty quantification is about thinking about the many alternatives, using the language of probability. For example, instead of: based on data, "parameter α is probably 4", "α is Gamma distributed with mean 4, variance 2".

UQ and Inverse Problems (continued)

- An important issue related to uncertainty quantification is resolution, particularly spatial/temporal in dynamics.
  - Higher resolution models: may be more computationally expensive to run; physics may not be well modeled.
  - Higher resolution observation more expensive to obtain.
  - Not clear if models have skill at high resolution.
- Statistical methods may be helpful to understand above.

Forward Model: Toy Example

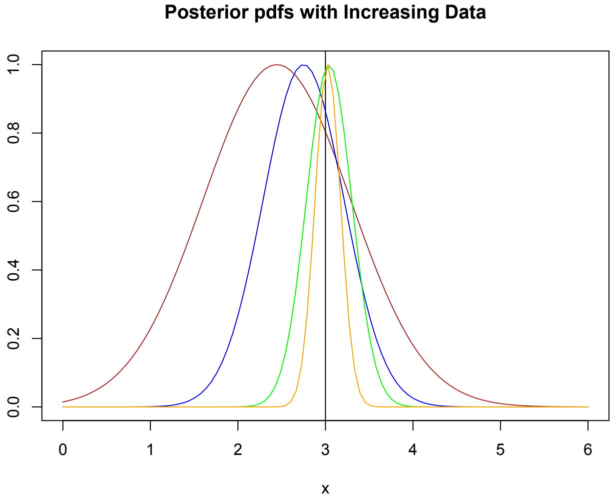
Forward model: $x_1, x_2, \ldots, x_n$ is Normal($\mu, \sigma^2 = 5$). μ is 3,5, or 2



Bayesian Inference: Toy Example

True forward model: Normal($\mu = 3, \sigma^2 = 5$)

Keep adding more data: Red $\rightarrow$ Blue $\rightarrow$ Green $\rightarrow$ Yellow



Why Use a Bayesian Framework?

- Lots of arguments about foundations, theory, philosophy, both for and against Bayesian inference.
- Some practical arguments in favor of Bayes in climate:
  - Often need to build models linking multiple data sets. Inference in such cases is generally simpler with Bayes. Natural way to build **hierarchical** or multi-level models.
  - More easily learn about relationships between parameters of interest, can "integrate out" nuisance parameters and examine important parameters marginally.
  - Natural framework to incorporate scientific knowledge; particularly important in ill-posed inverse problems.

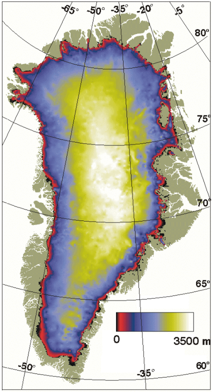
Bayesian Framework (continued)

- Results are immediately interpretable in terms of probabilities/probability distributions.

  Note: likelihood functions are *not* probability distributions and hence cannot be interpreted as such.
- Advantage of probability distributions is that predictions may be done seamlessly, while incorporating uncertainties. Algorithmically:
  - Draw multiple samples of parameters from posterior distribution.
  - For each parameter sample, simulate a draw from the probability (forward) model.
- Propogation of uncertainty is easy to do, that is: uncertainty about parameters $\rightarrow$ + uncertainty about model $\rightarrow$ + uncertainty about scenario $\rightarrow$ projections

Greenland Ice Sheet

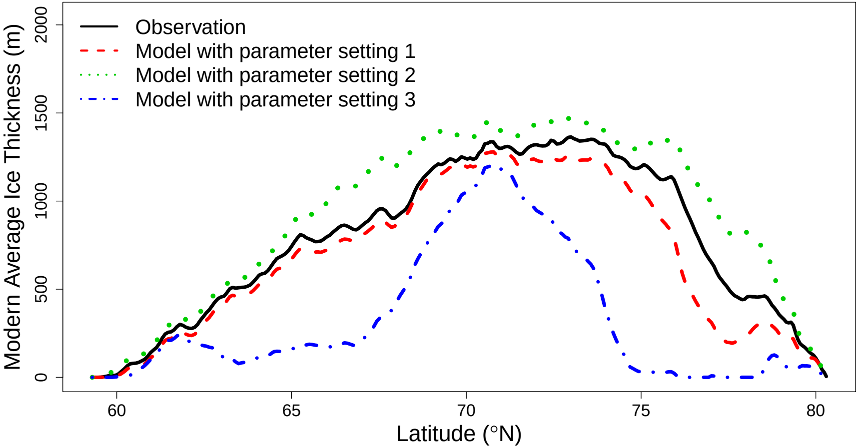
Important contributor to sea level rise: Total melting results in sea level rise of 7m.



Bamber et al. (2001)

Calibration Problem

Which parameter settings best match observations?



A Statistical Challenge in Climate Science

Focus here on one important challenge:

- Characterizing values for unknown or uncertain parameters (θ) of a climate model is called **calibration.**
- Informally: done by comparing climate model output (for various parameter values) to observations.
- Statistical model in two stages:
  1. Build an "emulator", $\eta(\theta)$ that captures relationship between θ and model output at θ, including for θ values at which model runs are not available
  2. Relate observations *Z* to the parameters $Z = \eta(\theta^*) + \delta(\theta^*) + \epsilon$, where δ is model-data discrepancy, ϵ is measurement error, θ^* is "fitted value" of parameter
- Can study δ and ϵ, sources of uncertainty
- Observations, model runs in the form of large spatial data

Statistical Methods

A Bayesian approach:

- **Prior distribution** plausibility of various parameter (θ) values: distribution $p(\theta)$
- **Probability model** (built using climate model runs) connects parameters to observations, accounting for model-data discrepancy.
- **Posterior distribution** plausibility of various values of the parameters *given* the data, integrating all the information and sources of uncertainty

Summary: Statistical models (Gaussian processes), data reduction (principal components), matrix theory (patterned covariances), inferential algorithms (Markov chain Monte Carlo)

Computing for Bayesian Inference

1. Statistical model is fit to $\mathbf{Y} = (Y(\theta_1), \ldots, Y(\theta_p))$.
   - Maximum likelihood: optimize parameters (ξ) of Gaussian process model likelihood function, $\mathcal{L}(\mathbf{Y}; \xi)$. Result: $\hat{\xi}$
2. Denote observations by Z. Obtain a probability model from above + discrepancy. Result: likelihood $\mathcal{L}_{\hat{\xi}}(\mathbf{Z}; \theta, \gamma)$
   - θ: calibration parameters, γ: discrepancy parameters
   - Bayesian inference based on posterior distribution, $\pi(\theta \mid \mathbf{Z}) \propto \mathcal{L}_{\hat{\xi}}(\mathbf{Z}; \theta, \gamma) p(\theta, \gamma)$,
   - Above conditional distribution contains all information about θ, γ, incorporating uncertainties
   - Markov chain Monte Carlo methods: sampling approach used to learn about complicated distribution $\pi(\theta, \gamma \mid \mathbf{Z})$.
   - Learn about climate model parameters θ and discrepancy.

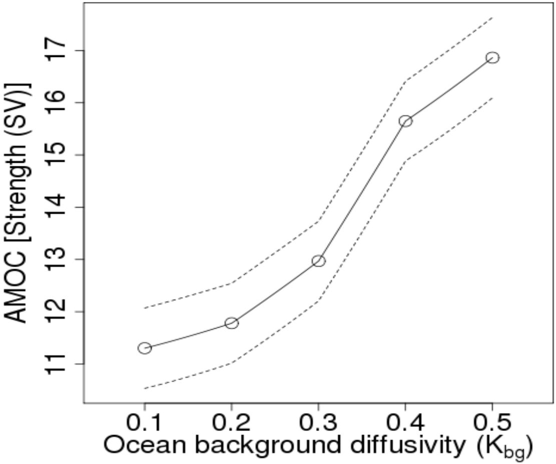
The AMOC and Climate Change

One concrete example:

- Atlantic Meridional Overturning Circulation (AMOC): AMOC heat transport makes a substantial contribution to the moderate climate of Europe (cf. Bryden et al., 2005)
- Any slowdown in the overturning circulation may have major implications for climate change
- AMOC projections from climate models.

A major source of uncertainty about the AMOC is due to uncertainty about $\mathbf{K_{bg}}$: model parameter that quantifies the intensity of vertical mixing in the ocean.

AMOC and Model Parameter K_{bg}

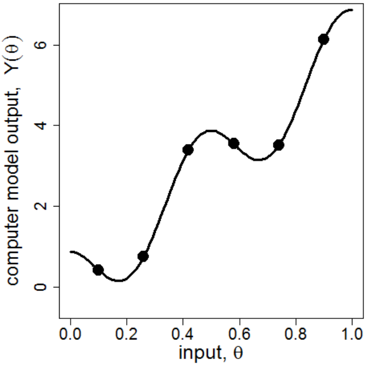


Learning about $\mathbf{K_{bg}}$

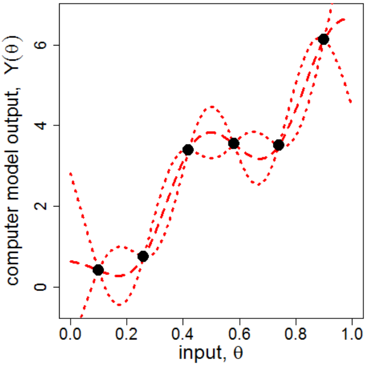
- Two sources of indirect information:
  - **Observations** of ocean temperatures.
  - **Climate model output** at different values of $\mathbf{K_{bg}}$ from University of Victoria (**UVic**) Earth System Climate Model (Weaver et. al., 2001).
- Models with different $\mathbf{K_{bg}}$ values result in markedly different ocean temperatures. Comparing observations to model output allows us to learn about $\mathbf{K_{bg}}$.

Emulation Step: A Simple Example

We use a statistical model called a **Gaussian process.** This model is a fast emulator (approximation) of the computer model.



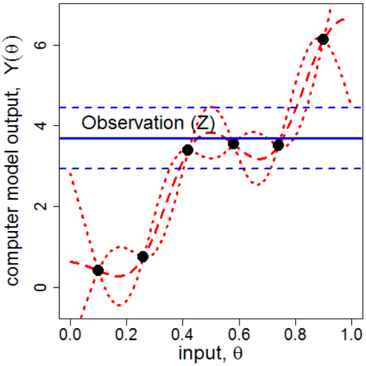
Computer model output (y-axis) vs. input (x-axis)



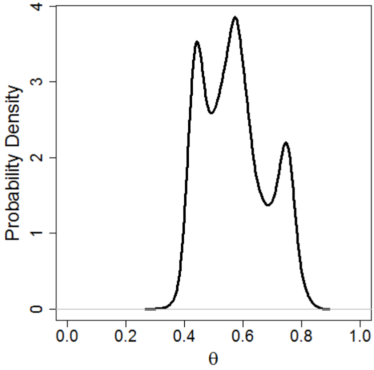
Emulation (approximation) of computer model using GP

Calibration Step: A Simple Example

We use statistical methods called **Bayesian inference and Markov chain Monte Carlo**: Use emulator (from before) and observations to learn about parameters.



Combining observation and emulator



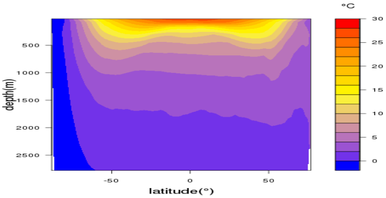
Posterior PDF of θ given model output and observation

Computational/technical challenges

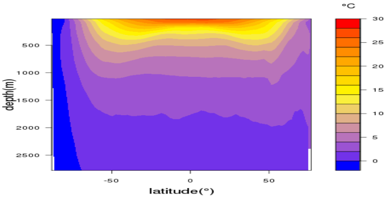
- We have 3D spatial observations+ climate model output.
- Using rigorous statistical methods can be prohibitively expensive for such data. Previous methods rely on aggregation. The effect of aggregation on uncertainties is not well understood.
- Example of SCRiM research contributions:
  1. New statistical methods and algorithms that allow us to work with the entire 3D data set without relying on aggregation.
  2. Comparison of results with unaggregated versus aggregated data: we can reduce uncertainties by using unaggregated data.

Ocean Temperatures

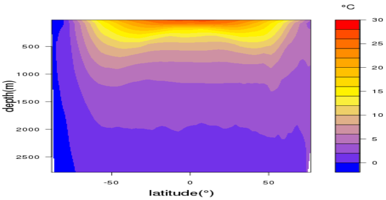
K_{bg} of 0.1



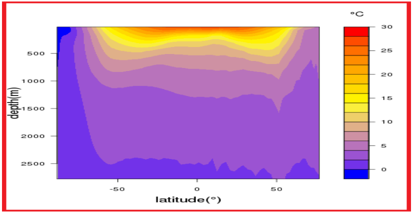
K_{bg} of 0.2



K_{bg} of 0.3

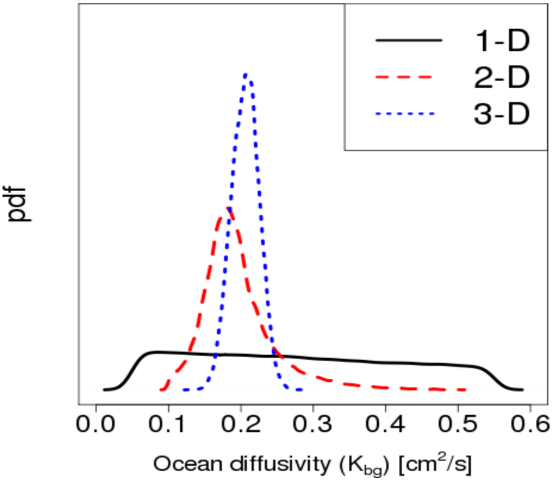


Observations



(2D versions of 3D data)

Results for K_{bg} Inference



(from Chang, Haran, Olson and Keller, 2013)

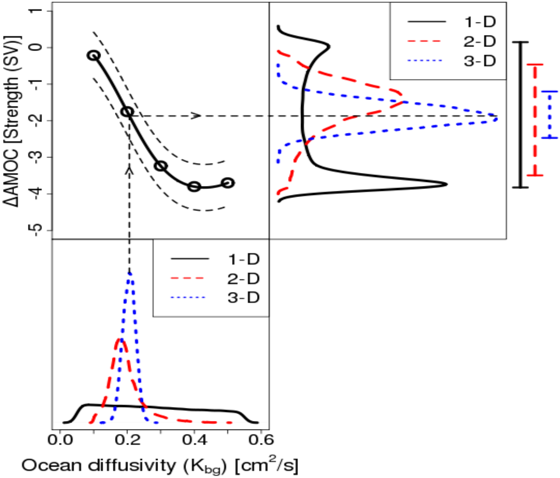
Results for $\mathbf{K_{bg}}$ Inference: Conclusions

- Our computationally efficient methods allow us to compare results from using aggregated (1D versus 2D) versus unaggregated (3D) data. Clear value:
  - Sharpest inference is based on unaggregated (3D) data.
  - Inference with 3D data is also robust to varying prior information; not so robust when using 2D or 1D data.
  - This results in sharper and more robust projections . . .
  - This is an example where UQ helps us answer questions related to the value of the resolution of models and observations.

Chang, Haran, Olson, Keller (2014, Annals of Applied Stats)

Related `R package` called stilt with simpler formulation

MOC Projections for 2100 Using Inferred K_{bg}

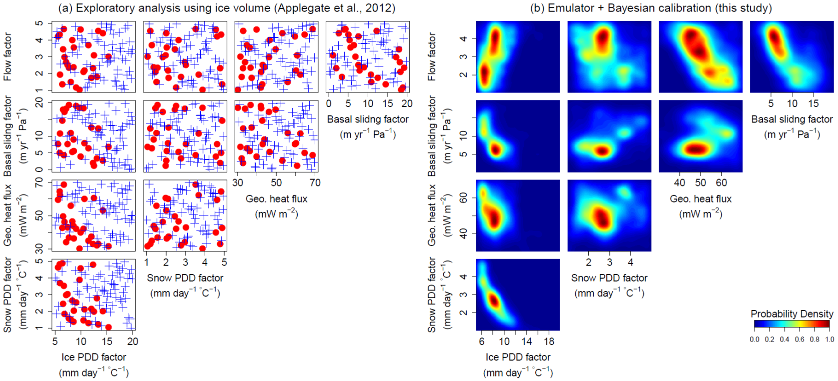


(from Chang, Haran, Olson and Keller, 2013)

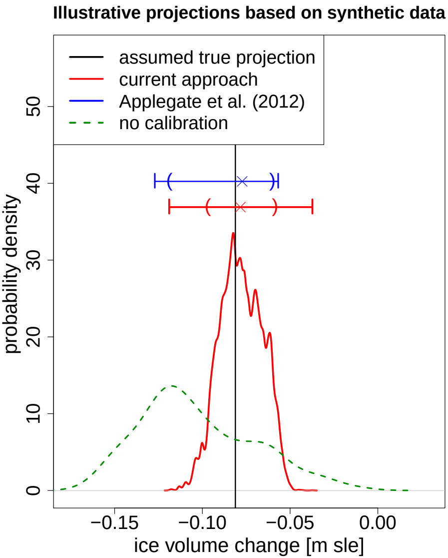
How Does Statistical Rigour Help?

Left: sensible but non-rigourous vs Right: sound statistics

"Underneath the hood": (i) accounting for (epistemic) uncertainties in emulation, (ii) real probability distributions.



Probability-Based Ice Volume Change Projection



Concluding Thoughts

- Without probability and statistics, it is not possible to quantify risk. Uncertainty quantification is central for science and policy
- Determining what the uncertainties are is a very difficult problem: error structures, dependencies, known unknowns, unknown unknowns . . .
- The mere act of discussing (debating) uncertainties is very useful. We can and should quantify many of those uncertainties. Of great interest for basic science.
- Can only be done if people work together (for a long time). Need collaborations between earth system scientists, statisticians, modelers, economists, . . .

Statistics Research Group

SCRiM Group: Interface of Earth System Analysis and Uncertainty Quantification

Meteorology, geosciences, statistics, Earth and Environmental Systems Institute (EESI)

- Lead Researchers: Chris Forest, Murali Haran, Klaus Keller, David Pollard, Rob Nicholas
- Grads/postdocs (former): Patrick Applegate, Won Chang, Yawen Guan
- Grads (current): Ben (Seiyon) Lee, Liz Eisenhauer

Relevant Manuscripts

- Chang, W., M. Haran, R. Olson, and K. Keller (2014): Fast dimension-reduced climate model calibration, *Annals of Applied Statistics*
- Chang, W., Applegate, P., Haran, M. and Keller, K. (2013) Probabilistic calibration of a Greenland Ice Sheet model using spatially-resolved synthetic observations: toward projections of ice mass loss with uncertainties
- Smith, L.A. and Stern, N. (2011) Uncertainty in science and its role in climate policy, Phil Transactions of the Royal Society A.
- Katz, R.W., Craigmile, P.F., Guttorp, P., Haran, M., Sanso, B. and Stein, M.L. (2013) Uncertainty Analysis in Climate Change Assessments, Nature Climate Change, 3, 769-771