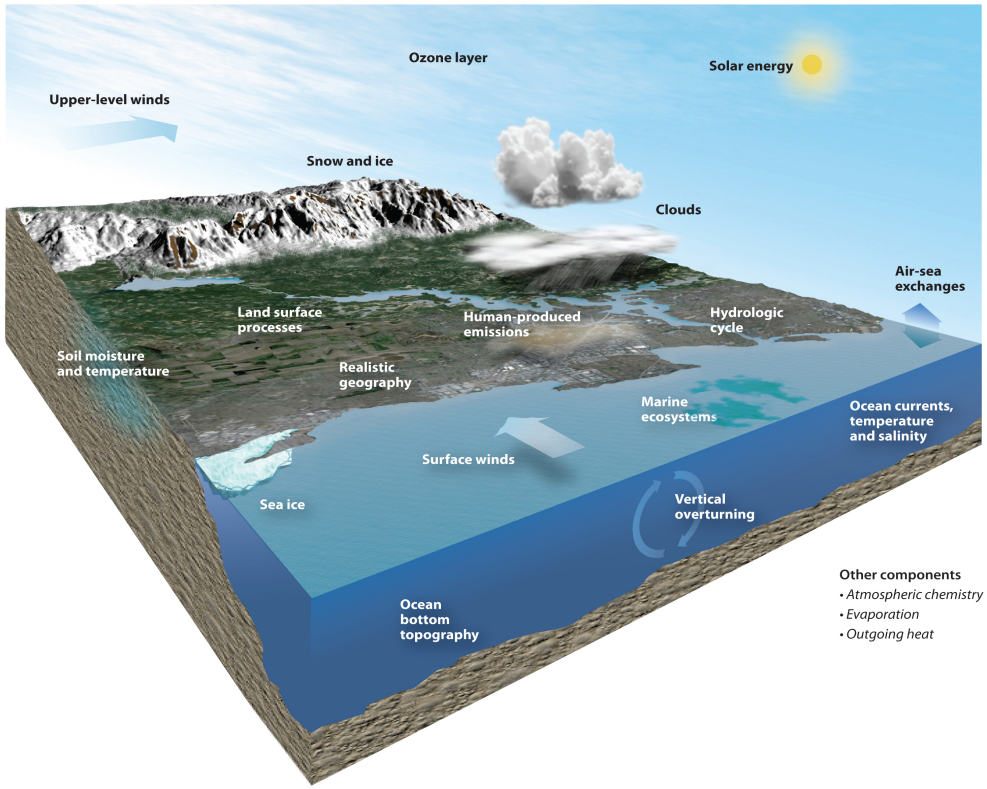
HECT: High-Dimensional Ensemble Consistency Testing for Climate Models

Nic Dalmasso[⋆1], Galen Vincent[⋆1], Dorit Hammerling[2], Ann B. Lee[1]

[1]Department of Statistics & Data Science, Carnegie Mellon University
[2]Department of Applied Mathematics and Statistics, Colorado School of Mines

NeurIPS: Tackling Climate Change with ML
December 11, 2020

Climate Models: NCAR Community Earth System Model[1]



- State-of-the-art "virtual laboratory" for studying past, present, and future global climate states;
- Fully coupled: all simulation components are computed together
- Code base currently contains > 1.5 million lines of code

[1] Image credit: `http://www.cesm.ucar.edu/`

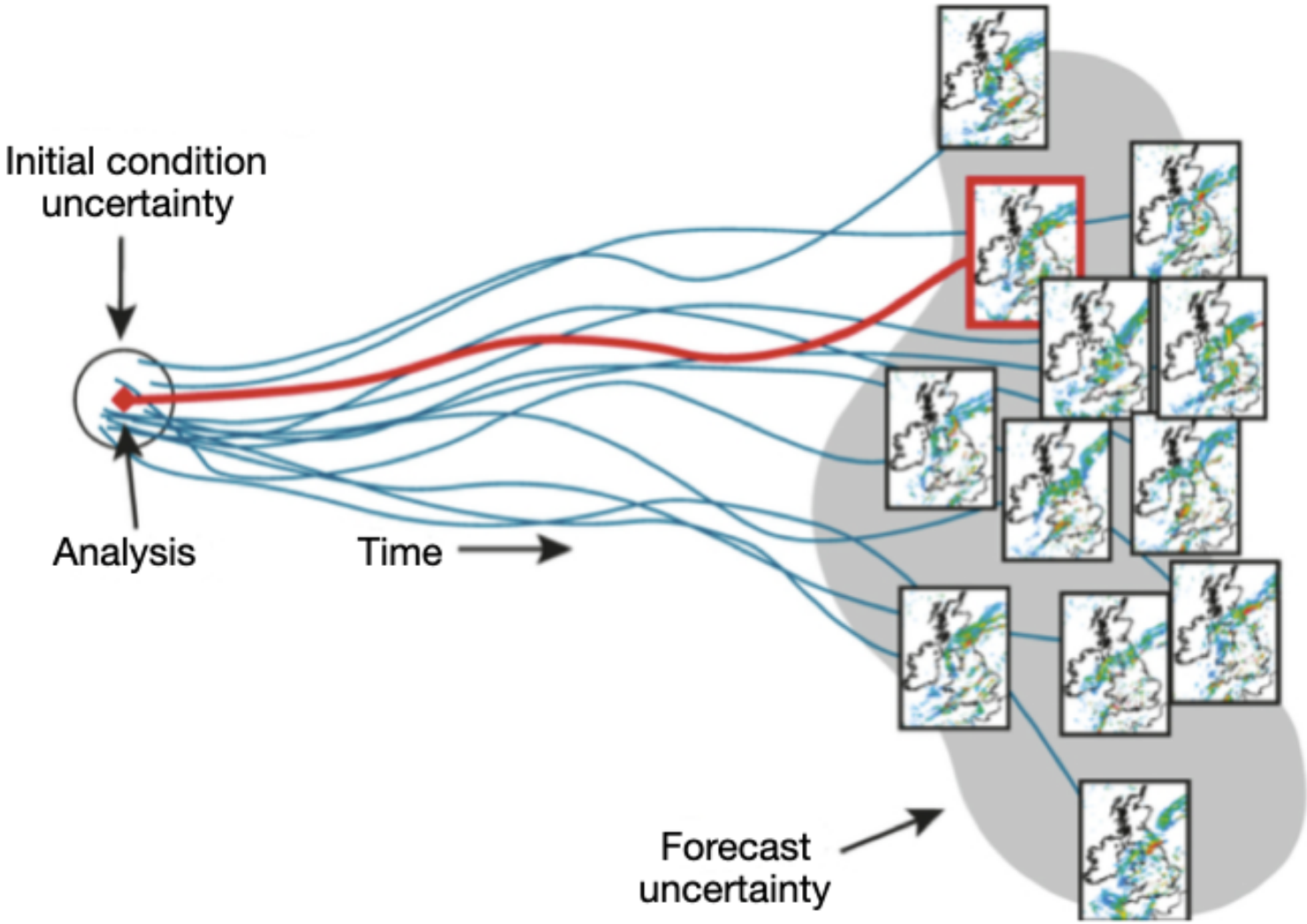
CESM is in constant state of development

- Quality assurance checks
  - Detect and reduce errors which could adversely affect the simulation
  - Maintain the model scientific credibility in the community

- Climate simulations can still be valid, but not bit-for-bit (BFB) identical to other runs
  - Different compiler or computing architecture
  - Different machine hardware
  - Different random number generator
  - Different parameter settings
  - ...

How can we make sure that non-BFB identical simulation outputs are a result of expected variation rather than a "climate-changing" error we introduced in the code?

Ensemble Consistency Testing



Compare the test simulation (red) with an ensemble of trusted simulations.

Two sample Test via Probabilistic Classifier

- Let P_0 and P_1 be the distributions of trusted and test runs
- Trusted runs: $\mathcal{S}_0 = \{\mathbf{X}_1^0, \ldots, \mathbf{X}_m^0\} \overset{i.i.d.}{\sim} P_0$
- Test runs: $\mathcal{S}_1 = \{\mathbf{X}_1^1, \ldots, \mathbf{X}_n^1\} \overset{i.i.d.}{\sim} P_1$ (usually $m \gg n$)

We want to formally test the hypothesis $H_0 : P_1 = P_0$ versus $H_1 : P_1 \neq P_0$

We turn this into a probabilistic classification problem:

- we introduce a binary random variable or class label Y for each run
- we interpret P_i, $i = 0, 1$, as the class-conditional distributions of $\mathbf{X}$ given $Y = i$

$$\implies H_0 : P_1 = P_0 \iff H_0 : \mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) = \mathbb{P}(Y = 1) \quad \forall\, \mathbf{x} \in \mathcal{X}.$$

HECT: High-dimensional Ensemble Consistency Testing

We can test such null hypothesis with the following test statistic[2]:

$$\widehat{\mathcal{T}} = \frac{1}{n+m} \sum_{i=1}^{n+m} \left(\widehat{r}(\mathbf{X}_i) - \widehat{\pi}_1\right)^2, \tag{1}$$

where $\widehat{r}(\mathbf{x})$ is an estimate of $\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$ based on $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{n+m}$, and $\widehat{\pi}_1 = \frac{1}{n+m} \sum_{i=1}^{n+m} I(Y_i = 1)$ is an estimate of $\mathbb{P}(Y = 1)$.

Benefits

- r can be any probabilistic classifier (from tree-based methods to deep neural networks – more later)
- Using high-capacity classifiers we can compare trusted and test runs at a spatial/temporal level (high-dimensional)
- We can provide diagnostics by identifying statistically significant spatial and/or temporal differences between runs

[2] Kim, Lee, Lei (2019), Dalmasso et al. (2020)

Comparison with State of the Art

Current approach – PCA-based testing:

- Compresses the data with a dimensionality reduction step
- Requires spatial and temporal averaging of simulation outputs, as well as climate variables selection
- Lacks theoretical guarantees for type I and II error

Proposed approach – HECT:

- No dimensionality reduction step is needed
- Spatial and temporal averages limited/not necessary
- Performance of the probabilistic classifier is shown to directly connect to the type I and type II error of the test[2]

Consistency Testing at Different Resolutions

Simulation outputs are currently compared after:

(1) Spatial averaging across a global grid and vertical atmosphere

(2) Temporal averaging across simulation time-steps

(3) Selection of relevant climate features

Examples of probabilistic classifiers for HECT:

(1) CNN can detect local differences in spatial structure

(2) Multivariate time-series models (e.g., RNN) can take into account the entire simulated sequence of climate variables

(3) Tree-based algorithms implicitly provide feature selection and are robust to highly correlated features

(1,2) Spatio-temporal deep neural networks can compare runs that are only averaged over the vertical level of the atmosphere

THANK YOU FOR READING!

A. H. Baker, D. M. Hammerling, M. N. Levy, H. Xu, J. M. Dennis, B. E. Eaton, J. Edwards, C. Hannay, S. A. Mickelson, R. B. Neale, D. Nychka, J. Shollenberger, J. Tribbia, M. Vertenstein, and D. Williamson.
A new ensemble-based consistency test for the community earth system model (pycect v1.0).
Geoscientific Model Development, 8(9):2829–2840, 2015.

Daniel J. Milroy, Allison H. Baker, Dorit M. Hammerling, and Elizabeth R. Jessup.
Nine time steps: ultra-fast statistical consistency testing of the community earth system model (pycect v3.0).
Geoscientific Model Development (Online), 11(2), 2 2018.

Ilmun Kim, Ann B. Lee, and Jing Lei.
Global and local two-sample tests via regression.
Electron. J. Statist., 13(2):5253–5305, 2019.

Niccolò Dalmasso, Ann Lee, Rafael Izbicki, Taylor Pospisil, Ilmun Kim, and Chieh-An Lin.
Validation of approximate likelihood and emulator models for computationally intensive simulations.
In *International Conference on Artificial Intelligence and Statistics*, pages 3349–3361. Proceedings of Machine Learning Research, 2020.