Developing Machine Learning Benchmarks for Weather and Climate Problems

David John Gagne
National Center for Atmospheric Research

June 9, 2022
Motivation

Big Question: How will we know when a machine learning model has made significant progress on a major climate challenge? When is an ML model ready to integrate into an Earth system model?

- What kind of benchmark is appropriate?
- How should we define our standard for “truth”?
- What kinds of metrics should be included in the intercomparisons?
- How can we make benchmarks more centered around end-users?
Types of Modeling Benchmarks

**Competition/Hackathon**
Measure relative performance on a specific task with a fixed dataset and metrics and provide users with practical experience on a given problem.

Examples: ImageNet, Kaggle Competitions

**Intercomparison**
Identify strengths and weaknesses of a set of models from different groups based on standardized procedures and suite of metrics.

Examples: CMIP

**Testbed**
Run models in a quasi-operational setting and receive subjective and objective feedback from end-users.

Examples: NOAA Hazardous Weather Testbed

**Readiness Tests**
Model must meet or exceed various safety, reliability, timing, and robustness criteria before being deployed into an operational setting.

Examples: Continuous Integration test suite
Organizers and Developers: Determines the problem, gathers the data, writes the code, creates and supports the platform.

Participants: Learn to use the data/models/software, submit new models, discuss issues, refine based on feedback, write papers, receive prizes,

Decisionmakers: Determine if new methods deserve further investment, perform holistic ranking of models, provide feedback to organizers and participants.
Benchmark Pathologies

- **Arbitrary Task Selection and Design**: tasks selected for benchmark based on convenience and ease of measurement even if not fully representative of problem
- **Over-generalization of benchmark scope**: achieving high performance on benchmark falsely equated with solving the problem
- **De-contextualized data and performance metrics**: people using benchmark data and metrics without original context
- **Subjectivity and cultural biases**: benchmark data encodes subjective beliefs of creators through data selection, labeling and quality control (or lack thereof)
- **Redirection of focus for a scientific field**: Strong benchmark performance can lead researchers to over-focus on top-performing methods at the expense of others

How can we avoid these pathologies in benchmark design?

Citation: Raji, I. D., E. M. Bender, A. Paullada, E. Denton, and A. Hanna, 2021: AI and the Everything in the Whole Wide World Benchmark. arXiv [cs.LG].
Components of Forecast Goodness

**Quality**
Correspondence between forecasts and observations

**Consistency**
Similarity of forecast distribution properties to physical system

**Value**
How user incrementally benefits from a better forecast

---

## Selecting “Truth” for a Benchmark

<table>
<thead>
<tr>
<th>Benefits</th>
<th>In-Situ Observations</th>
<th>Remote Sensing</th>
<th>Analysis</th>
<th>Model Output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Closest record of what actually happened</td>
<td>Greater coverage in 4D compared with in-situ obs</td>
<td>Optimal combination of observations and model output</td>
<td>Self-consistent state variables available across uniform grid and over any time period when model can be run</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data on uniform grid and regular timesteps</td>
<td>Can be configured for arbitrary stationarity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Estimate uncertainty</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Perils</th>
<th>In-Situ Observations</th>
<th>Remote Sensing</th>
<th>Analysis</th>
<th>Model Output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sparse in time and space</td>
<td>Gaps in coverage due to type of instrument and sampling strategy</td>
<td>Includes biases from both observations and models</td>
<td>Includes model biases</td>
</tr>
<tr>
<td></td>
<td>Non-stationarity in observing system availability and standards</td>
<td>Requires physical/statistical retrieval procedure to derive variables of interest</td>
<td>Non-stationary in time because of changes in observing systems</td>
<td>May be missing key processes found in observations</td>
</tr>
<tr>
<td></td>
<td>Distributed preferentially in certain locations and minimal in others</td>
<td>Non-stationarity in space and time</td>
<td>Sensitive to interpolation procedure</td>
<td>Can be expensive to generate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Can be less noisy than observation data</td>
</tr>
</tbody>
</table>
CAM vs. DA Modes of Variability

Meridional Surface Wind

V EOF number 0

DA:

Nudge:

Correlation

Slide courtesy Will Chapman
Problem: Using a simple random train/test split on space/time-correlated data
Fix: Group examples by sufficiently long time ranges to minimize cross-correlation and split by group

Problem: Re-tuning your model settings after seeing test set results
Fix: Tune hyperparameters of model with validation test and withhold test set until final evaluation. Test final model on similar but independent data or real-time data.

Problem: Pre-processing transformations include testing data in calculations
Fix: Treat pre-processing as part of model and only use training data for fitting transformations
Types of Consistency

Spatial Consistency: ensuring forecast variation matches spatial variation and properties

Types of Consistency

Explanation Consistency: how closely do equivalent models trained on same data agree on their importance rankings and sensitivities?


Gagne, Becker, Schreck, Gantos, Molina, Sha, Quantification of XAI Model Uncertainty, In Prep.
Climatological Consistency

Figure by Charlie Becker
Value

Value is the incremental benefit (personal, monetary, societal) to user outcomes realized by using a forecast or projection.

Value for a user is realized through the chain of processes required to make and disseminate the forecast. If chain is broken, value cannot be realized.

What can add value to users beyond better quality and consistency?

- Faster time to solution
- Reduced computational requirements
- Ease of development
- Ease of use
- Effective visualization

https://hwt.nssl.noaa.gov/sfe_viewer/2022/outlook_verification/
**User role description:** Create personas about end-users of a given challenge problem algorithm

**Department of Transportation Official**

**Background:** Wind speeds are an important factor for transportation officials when making decisions about closing and (reopening) bridges. Strong winds can cause driving conditions to deteriorate, making driving on them incredibly dangerous. Crashes are not only dangerous to those driving, but they could also block important routes for emergency response vehicles. The decision to close the bridges is also very important because they are needed for people to evacuate, especially from barrier islands, so closing the bridge means potentially taking away some people's ability to evacuate from a storm.

**Key user needs:** The transportation official needs extremely precise data on wind speeds so they can effectively walk the line between keeping pathways open for evacuation and making sure driving conditions are safe.

---

**Trustworthy AI Approach Evaluation:**
Participants perform different kinds of ML evaluations on pre-trained models each day
- Verification statistics
- XAI (shallow and deep)
- Case studies and failure modes
- Uncertainty Quantification

---

**Daily Value Reflection**
How does each kind of ML analysis provide insight (if any) about the value the ML provides for the user persona?
Aim to build community and cooperation, not competition. Benchmarks are a community effort, so design for all participants getting positive feedback from experience.

Make the benchmarking process accessible to all audiences

Organizers: open development on shared platforms

Participants: moderate data sizes, provide compute, accessible libraries and data formats

Decision makers: Produce summary papers and websites with in-depth insights

Center benchmarking around synchronous events. Events help share context and build momentum for groups to work together. Purely asynchronous benchmarks can encourage metric-chasing and loss of context.
Prompt 1: Major uncertainties identified in IPCC AR6 (grand challenge) and potential to reduce it with your method: New ML Climate Benchmarks could provide a standardized way to evaluate whether ML-climate model integrations actually decrease uncertainty.

Prompt 2: Impact of your method/model on a 5-year time-scale (roughly CMIP7): Improved benchmarks will enable holistic evaluation of new ML approaches to help working groups decide whether to move forward with ML components in CMIP 7 models.

Prompt 3: Potential benefits for robust, trustworthy, and actionable climate measurements and models: Incorporating a suite of quality, consistency, and value metrics will give a more complete picture of goodness.

Prompt 4: Generalizability and extensibility to other climate problems: Many evaluation methods are generalizable across climate problems (assuming availability of reasonable truth, consistency expectations, and users).

Prompt 5: How to engage the community: Co-design of benchmarks with community is critical to ensuring the benchmarks are useful. Benchmarks can also help foster new communities.

Prompt 6: List the main challenges: Avoiding benchmark pathologies. Truth may not be readily available. Keeping people engaged with benchmark over longer periods of time.