FY23 DoD Cloud Workshop

Cube of Calibrated Clouds

PI: Justin McLay
justin.mclay@nrlmry.navy.mil
Dan Hodyss, NRL Remote Sensing Div.
Matthew Fernandez, Bay Systems Inc.
Glen Carl, SAIC Inc.

13 September 2023

Cube of Calibrated Clouds for the NAVGEM Ensemble

Short History of Navy Global Ensemble Post-Processing

Timeline of Navy global ensemble post-processing projects:

1) FY16-FY17 ONR 6.4 National Unified Operational Prediction Capability (NUOPC)
   - Bias correction of significant wave height

2) FY18-19 ONR 6.4 Forecast Improvement through Statistical Post-Processing
   - Bias correction of near-surface winds

3) FY20-FY22 ONR 6.4 Enhanced Probability Distributions
   - Bias correction and calibration of certain variables for naval aviation UAV ISR
     i. Cruise-level air temperature and wind speed
     ii. Probability of icing severity
     iii. Probability of turbulence severity

4) FY23- ONR Cube of Calibrated Clouds from the NAVGEM Ensemble
   - Began December 2022
   - Bias correction and calibration of cloud variables

- FY18-21 ONR 6.4 Triton MQ-4C Hazardous Weather Avoidance

- FY19-21 NRL 6.2 Satellite-Model Synergies for Hazardous Event Nowcasting
- FY22- Joint Depiction of Atmospheric Clouds (JDAC) (Nachamkin)
- FY22- Unified Cloud Regime Verification (Christophersen)

Cube of Calibrated Clouds for the NAVGEM Ensemble

Overview

Technical Capability

A "calibration" system that generates a 4D data cube of unbiased and calibrated probabilistic cloud forecasts from the surface to top-of-atmosphere and across the short-range (less than one week) forecast interval.

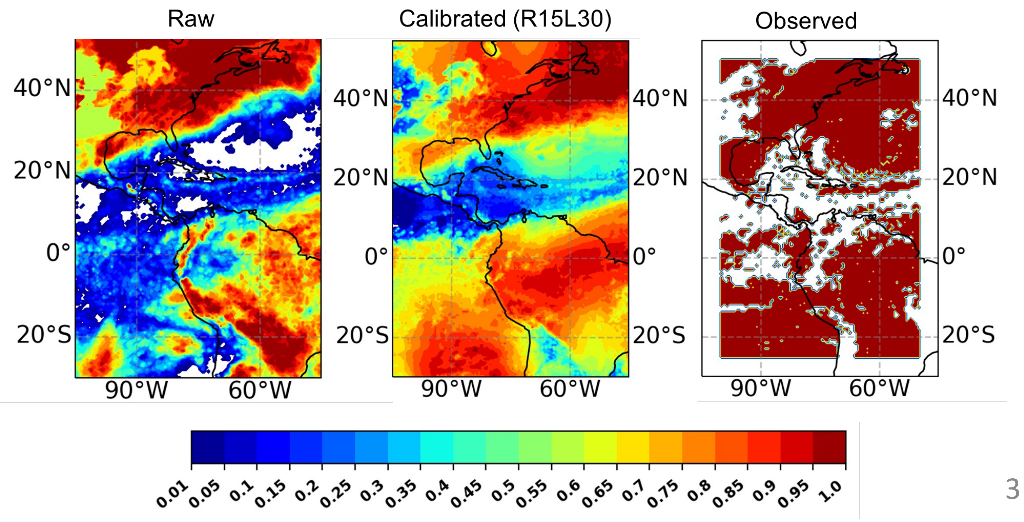
The calibration uses the historical relationship between the forecast and observations to adjust the mean and variance (and other moments) of today's forecast ensemble.

FY22-23 Accomplishments

- We successfully updated our existing post-processing code base to handle cloud variables.
- We completed a benchmark test case using the updated code to post-process probability of total cloud cover, with very good validation results.
- We completed the hiring of a new data scientist.

Operational Impact

- There is potential for broad operational impact: This work can seamlessly benefit any operation or downstream system that relies on global cloud data.
- Many research applications in 6.1 and 6.2 projects.
- Long-term potential for further performance gains
  - We are just starting to consider the more sophisticated calibration methods.



Milestones and Work Unit Outline

The project proposes a two-part (basic/advanced) approach to cloud calibration:

- **Benchmark solution**: Uses traditional statistical methods, and builds off the existing post-processing software package.
- **Advanced solution**: Uses some form of ML/AI, to be determined by training data availability and the nature of the cloud forecast problem.

- FY22-23:
  - Milestone 1:

    Coordinate with Jason Nachamkin to leverage the JDAC verification apparatus
  - Milestone 2:

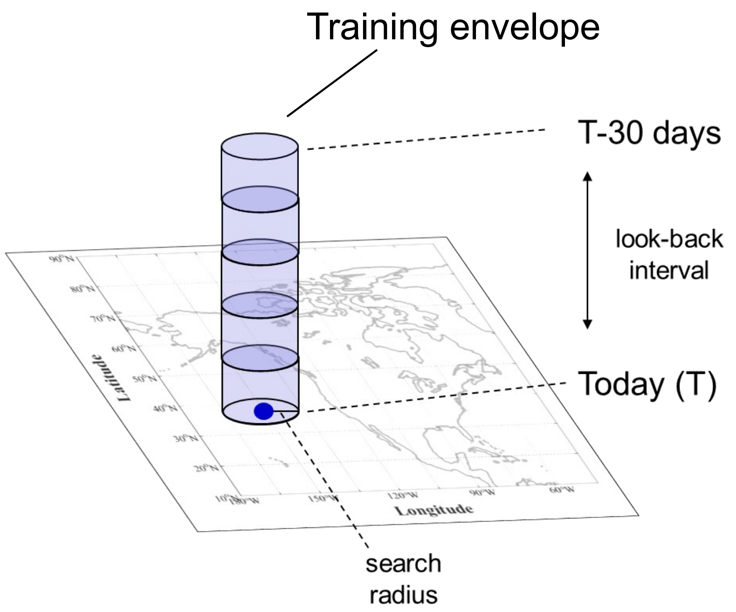
    Build archive of the necessary NAVGEM ensemble fields
  - **Milestone 3:**

    Develop a benchmark calibration for total cloud cover/type based on traditional statistical methods (by adapting the existing post-processing system)
  - Milestone 4:

    Begin groundwork for the advanced calibration based on ML/AI

Procedural Details of the Benchmarking

- Methods: linear regression, quantile mapping (e.g. Hamill and Scheuerer 2018)
- Variable: column total cloud fraction
- Event: sky cover is broken or overcast, e.g. cloud fraction > 0.5
- Test period: January-February 2023 (out-of-sample test)
- Resolution: 720x361 (0.5-degree) fields
- Validation: GOES16E observations (Atlantic basin)



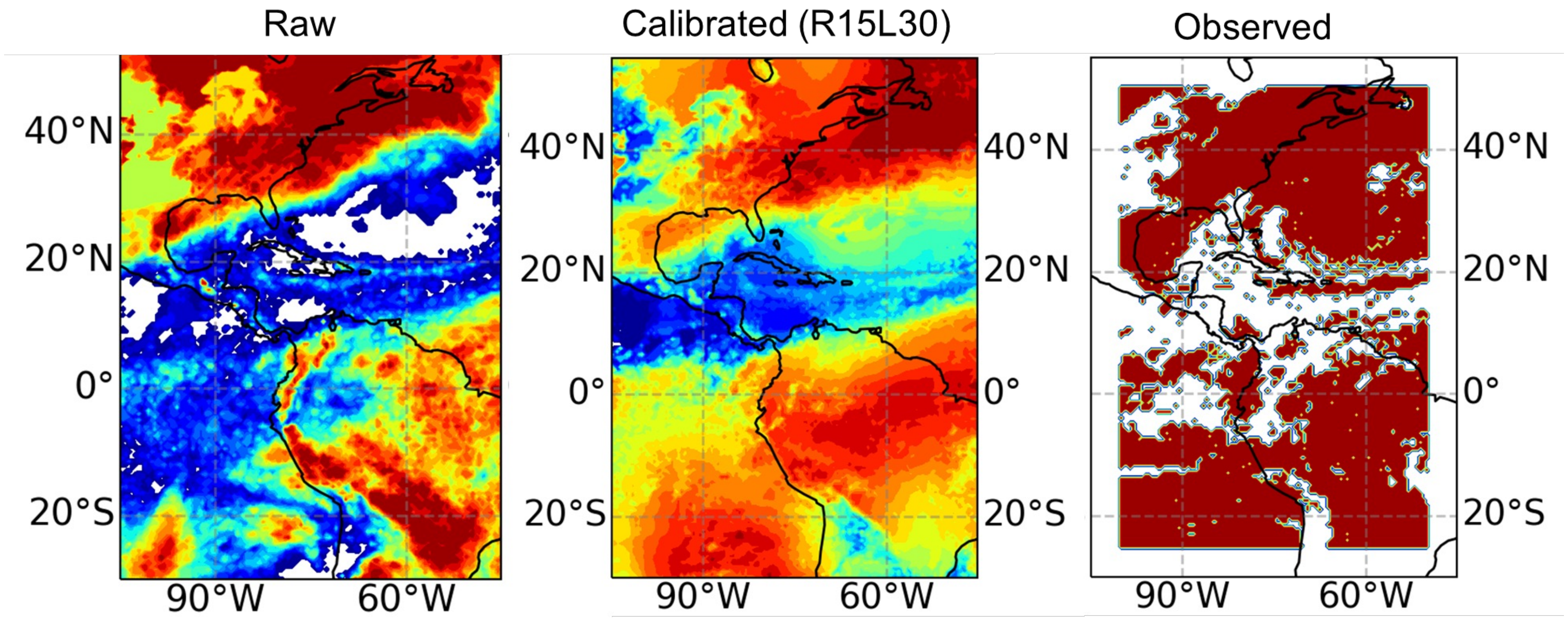
- The post-processing is solved locally, grid point by grid point
  - Training data envelope governed by two parameters:
    i. Spatial search radius: e.g. 5, 10 deg.
    ii. Temporal look-back interval: e.g. 10, 30, and 60 days.

- Computational resources: Distributed computing, typically ~16 CPUs

- Validation metrics:
  - Reliability diagrams
  - Brier scores
  - Receiver operating characteristic (ROC)
  - CRPS

Early Benchmark Cloud Calibration Results

Case study of calibration for 2023021712:



(Left and center) T+48h ensemble forecast probability of total cloud cover being in the broken or overcast category, initialized 2017021512 and valid 2023021712.

(Right) Observed probability valid 2023021712.

The calibration enhances the probability of broken or overcast conditions across broad swaths of the domain, including over the subtropical Atlantic and South Pacific, improving agreement with observations.

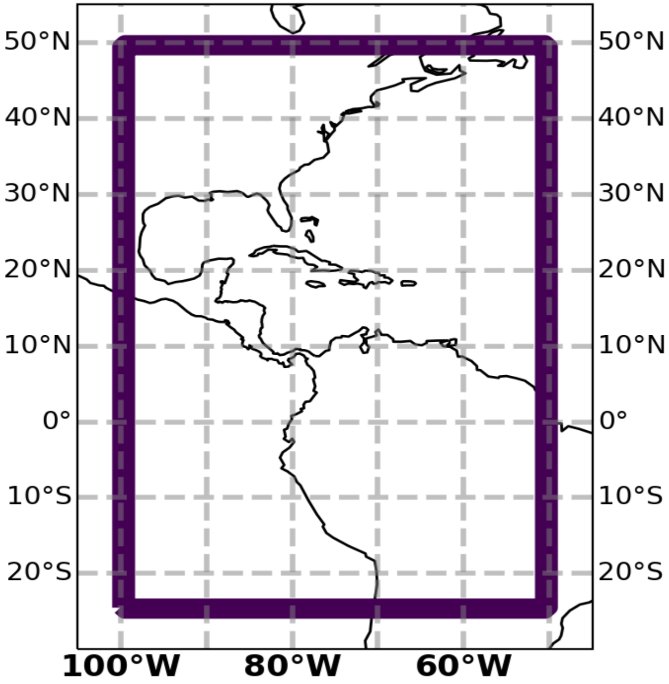
Early Benchmark Cloud Calibration Results

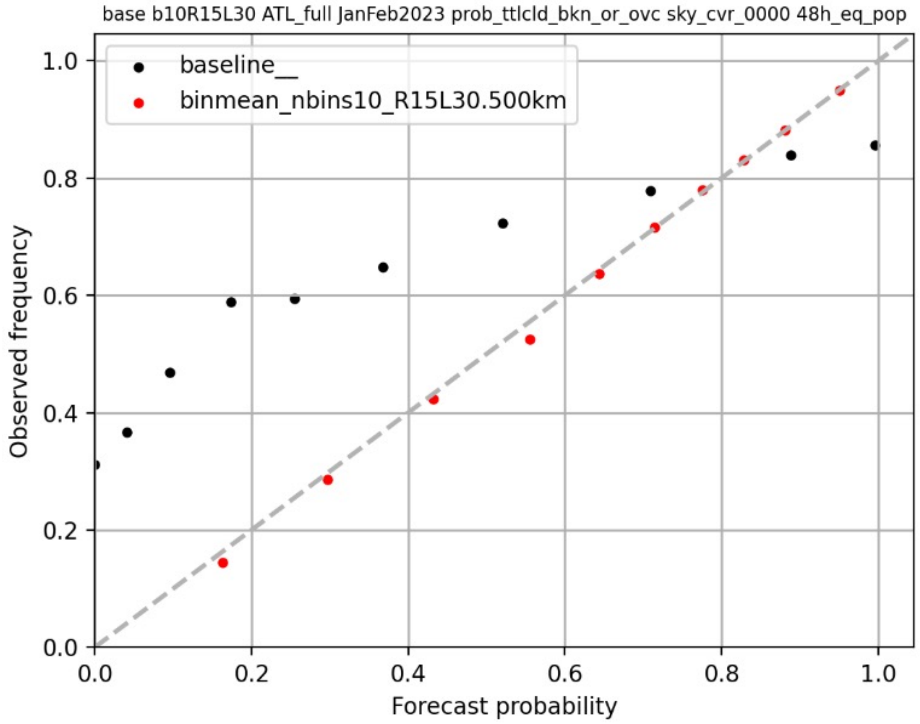
Diagnostics of forecast reliability:

(Left) Validation domain over the Atlantic basin.

(Right) Reliability diagram for the T+48h ensemble forecast probability of total cloud cover being in the broken or overcast categories, constructed across forecasts initialized on 28 DTGs in February 2023.

The raw (calibrated) forecast probabilities are indicated by the black (red) filled circles.





The data points corresponding to the raw forecast tend to fall well above the ideal 45-degree line of the plot, indicating that the raw forecast strongly under-predicts the likelihood of broken or overcast conditions. Meanwhile, the points corresponding to the calibrated forecast align very closely with the ideal 45-degree line, indicating much improved agreement with observations

Cube of Calibrated Clouds for the NAVGEM Ensemble

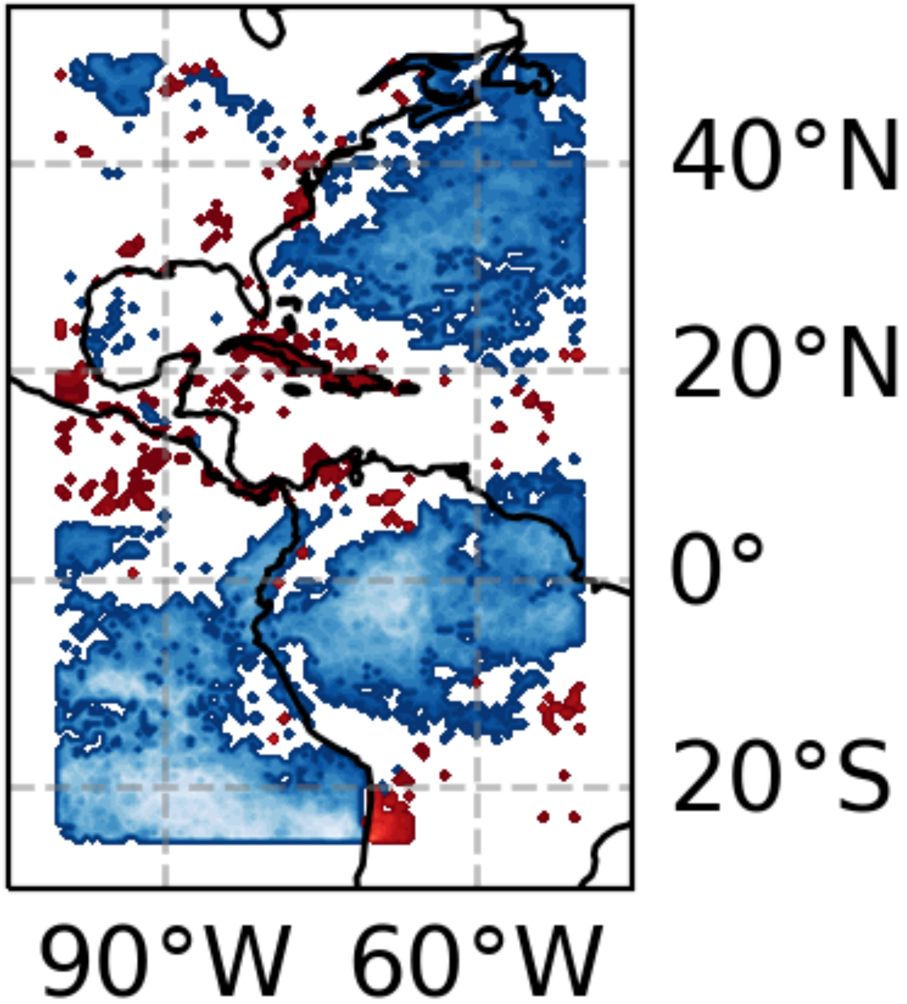
Early Benchmark Cloud Calibration Results

Diagnostics of forecast Brier score:

(Right) Brier score for the T+48h ensemble forecast probability of total cloud cover being in the broken or overcast categories, constructed across forecasts initialized on 28 DTGs in February 2023.

Blue (red) shading indicates that the Brier score of the calibrated forecast is significantly better (worse) than the score of the raw forecast.

This also includes relative improvement (RI) criterion of 30.0%



Cube of Calibrated Clouds for the NAVGEM Ensemble

Next Steps After Benchmarking: Methodological Precedents

Some influential methodological precedents for the planned advanced solution (not an exhaustive list)

- Traditional
  - Model Output Statistics (MOS) (Glahn and Lowry 1972)
  - Nonhomogeneous Gaussian regression (EMOS) (Gneiting et al. 2005)
  - Logistic regression
    - Hemri et al. 2016: Discrete Postprocessing of Total Cloud Cover Ensemble Forecasts
  - Bayesian model averaging (BMA) (Raftery et al. 2005)
- ML/AI
  - Random forests (Baran et al. 2021: Machine learning for total cloud cover prediction)
  - Neural network approaches:
    - Baran et al. 2021: Machine learning for total cloud cover prediction
      - Multilayer perceptron (a traditional network with non-convolutional structure)
    - Dupuy et al. 2021: ARPEGE Cloud Cover Forecast Postprocessing with Convolutional Neural Network (U-Net)
    - Dai and Hemri 2021: Spatially Coherent Postprocessing of Cloud Cover Ensemble Forecasts
      - Conditional generative adversarial network (cGAN)
    - Agrawal et al. (Google): Convolutional Neural Network for predicting cloud mixing ratio
    - WMO S2S AI Challenge (not specific to clouds): Landry et al. Convolutional neural network

Cube of Calibrated Clouds for the NAVGEM Ensemble

Our Realities Regarding Training Data for Navy Ensemble Clouds

- Neural network approaches tend to require high quality, curated, large datasets for training (e.g. Haupt et al. 2021)
- The Navy OPS global ensemble:
  - Output files are not archived long-term.
  - Only limited number of forecast fields are output.
    - Total cloud fraction, but no low-, mid-, or high-cloud fraction
    - Lacking fields required to discriminate cloud cover type (convective versus stratiform), e.g. convective accumulated precipitation, cloud base/top, …
- In principle, we might be able to back-out forecast cloud cover at various vertical levels given the forecast basic state variables (T, RH, …).
  - There might be hidden catches/difficulties with this.
- Preferred option: Generate in-house multi-year NAVGEM ensemble reforecast dataset. Proposed for FY24.
  - Gives us full control over output and archiving.
  - At the current OPS ensemble resolution of T359L60, not that expensive to run 1x/week for 15 years.
- In absence of a reforecast dataset, some options from the ML/AI community are:
  - Transfer learning (e.g., using ECMWF or NCEP reforecasts as proxy training data).
  - Data augmentation
- Last ditch option: revert to traditional methods, possibly just w/ total cloud fraction, no discrimination of cloud type

Significance and Navy Relevance

The project will achieve a number of important milestones for Navy global ensemble forecasting:

- **It will establish the Navy's first ever ensemble calibration system for cloud variables of any sort (cloud cover/type).**
- It will provide improved probability forecast skill for total cloud cover (type) over the [T+0h, T+144h] lead time interval. Based on the post-processing results of 6.4 Enhanced Probability Distributions and the predecessor projects, the gains in skill may be substantial.
- It will expand the portfolio of post-processed fields for the NAVGEM ensemble, and continue with the emphasis on the post-processing of operationally important variables (icing severity, turbulence severity, cloud cover/type) that was established under 6.4 Enhanced Probability Distributions and the predecessor projects.
- **We will learn what is wrong with the raw cloud forecasts from the NAVGEM ensemble, e.g., what the biases are, what regions have over- or under-predicted cloud probabilities, etc.**
- We will learn what traditional post-processing approaches can do in terms of correcting errors in global cloud forecasts, and what they can't do. I.e., what types of errors can be corrected by these approaches, and what types need further research.
- **Similarly, we will learn what ML/AI approaches can, and can't, do in terms of correcting errors in global cloud forecasts.**

Challenges and Risks

Presently we see two long-term moderate risks to the project:

1) One is our ability to gather sufficient training data to effectively employ a calibration analytic based on one of the more sophisticated ML/AI methods, e.g. a deep convolutional neural network (CNN).

    These methods tend to require a large amount of training data. We have strategized alternatives for obtaining sufficient data, but there is no guarantee these will be satisfactory.

2) Another is our ability to find adequate HPC resources for running the advanced ML/AI methods.

    Experience has shown that the HPC machines with GPUs that are preferred for ML/AI computation see intense use and often have relatively poor job throughput. We are considering alternatives that include:

    running our ML/AI applications at CIRA

    purchasing our own in-house GPU cluster.

    However, these alternatives may not be viable.