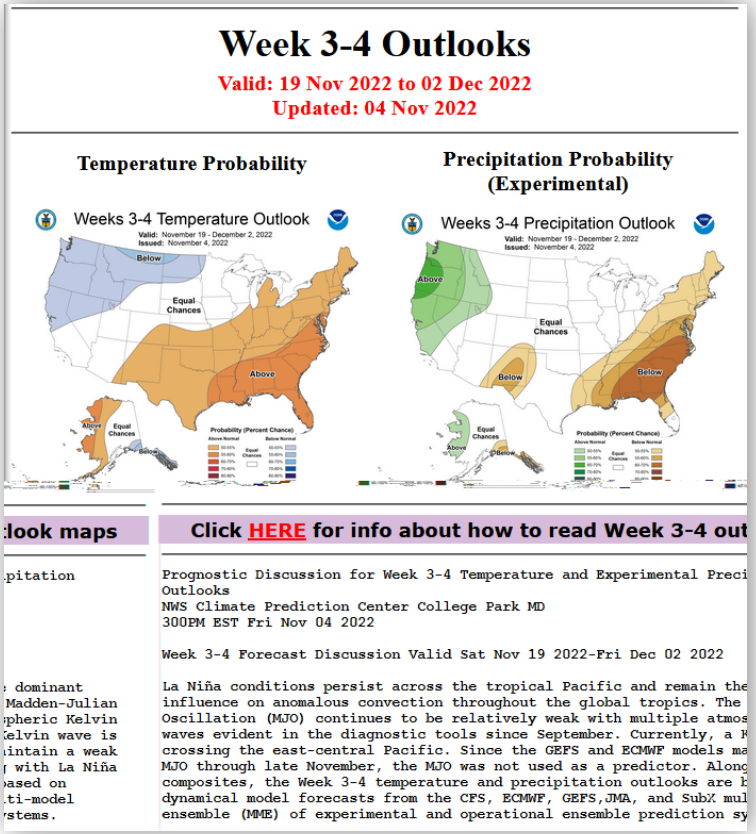
Bayesian Joint Probability (BJP) Calibration of Subseasonal Model Forecasts

Dan C. Collins, Johnna M. Infanti[1]
Sarah Strazzo[2]
QJ Wang[3]
Andrew Schepen[4]

[1]NOAA/NWS/CPC
[2]Embry-Riddle Aeronautical University
[3]University of Melbourne
[4]CSIRO

Week 3-4 Forecasts at CPC



Week 3-4 Outlooks

Valid: 19 Nov 2022 to 02 Dec 2022
Updated: 04 Nov 2022

Temperature Probability

Weeks 3-4 Temperature Outlook

Valid: November 19 - December 2, 2022
Issued: November 4, 2022

Below
Equal Chances
Above

Probability (Percent Chance)

Precipitation Probability (Experimental)

Weeks 3-4 Precipitation Outlook

Valid: November 19 - December 2, 2022
Issued: November 4, 2022

Above
Equal Chances
Below

Probability (Percent Chance)

look maps

Click HERE for info about how to read Week 3-4 out

pitation

dominant
Madden-Julian
pheric Kelvin
Kelvin wave is
aintain a weak
with La Niña
ased on
ti-model
stems.

Prognostic Discussion for Week 3-4 Temperature and Experimental Preci
Outlooks
NWS Climate Prediction Center College Park MD
300PM EST Fri Nov 04 2022

Week 3-4 Forecast Discussion Valid Sat Nov 19 2022-Fri Dec 02 2022

La Niña conditions persist across the tropical Pacific and remain the
influence on anomalous convection throughout the global tropics. The
Oscillation (MJO) continues to be relatively weak with multiple atmos
waves evident in the diagnostic tools since September. Currently, a K
crossing the east-central Pacific. Since the GEFS and ECMWF models ma
MJO through late November, the MJO was not used as a predictor. Along
composites, the Week 3-4 temperature and precipitation outlooks are b
dynamical model forecasts from the CFS, ECMWF, GEFS,JMA, and SubX mul
ensemble (MME) of experimental and operational ensemble prediction sy

CPC releases week 3-4 outlooks each week on Friday
https://www.cpc.ncep.noaa.gov/products/predictions/WK34/

- 2-Category probabilistic forecasts are provided for temperature and precipitation
- Temperature probabilities are calculated with respect to above/below the mean
- Precipitation probabilities are calculated with respect to above/below the median
- Maps are based on *multiple* tools, with the forecaster creating the official outlook using information from dynamical models, statistical models, and climate state indices such as MJO, etc.

A prognostic discussion included with the forecast provides information about what tools were considered, important climate drivers, and areas of greater/lesser certainty.

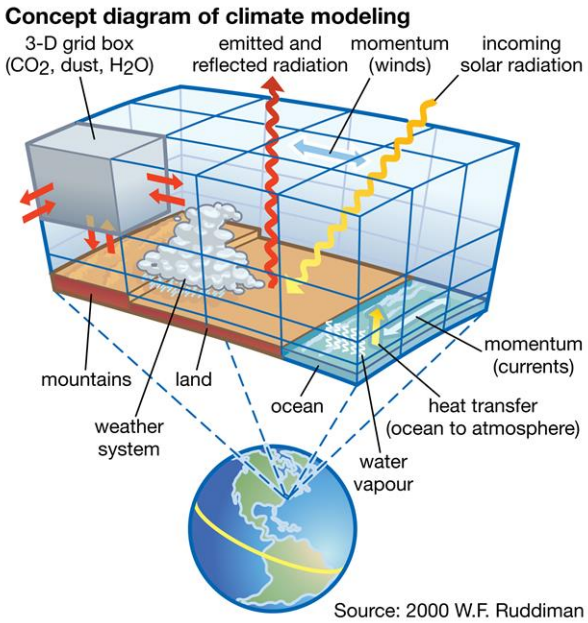
Model Calibration - A Crash Course

What is calibration?

Statistical post-processing of dynamical climate model output with the goal of improving skill and reliability

Why calibrate?

Dynamical models are skillful at simulating large-scale climate features, but are generally biased due to systematic errors and statistically unreliable in determining likelihood

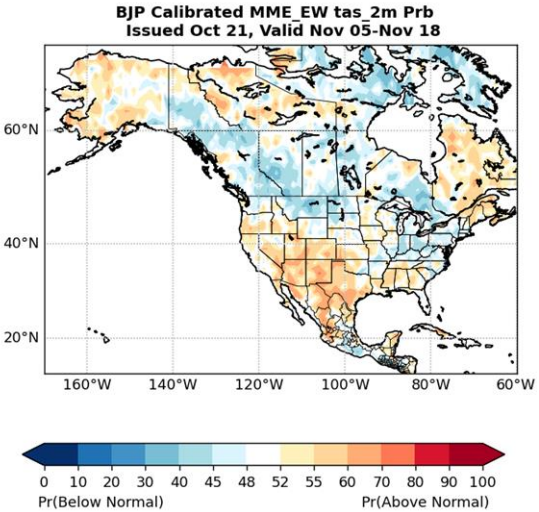


A central problem of dynamical models is parameterization of physical processes. Forecasts can be improved through advances in model physics parameterizations or increased resolution

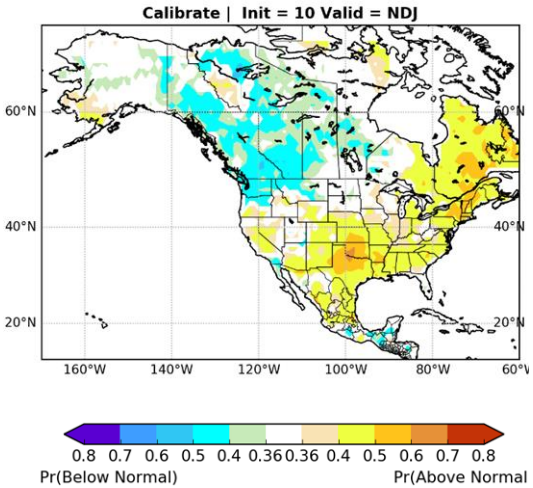
Or through statistical calibration!

Bayesian Joint Probability (BJP) Calibration

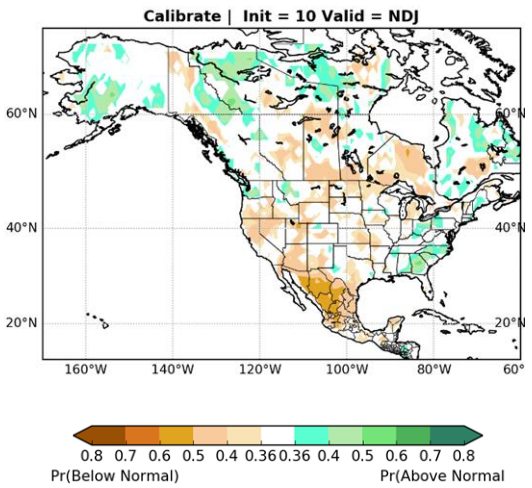
In partnership with CSIRO scientists and as a result of two prior funded proposals, CPC has integrated a calibration methodology into experimental subseasonal and seasonal prediction tools using dynamical models - This method is **Bayesian Joint Probability** (BJP)



Subseasonal BJP Calibration (3 model SubX MME). Currently 2 category temperature.



Seasonal BJP Calibration (NMME). Tercile temperatures



Seasonal BJP Calibration (NMME). Tercile precipitation.

Overview: BJP Calibration

- **Calibration:** Relates dynamical model output to observed climate variables in order to correct both model bias and ensemble spread. Forecasts are returned to climatology in the absence of correlation between forecasts and observations.

Overview: BJP Calibration

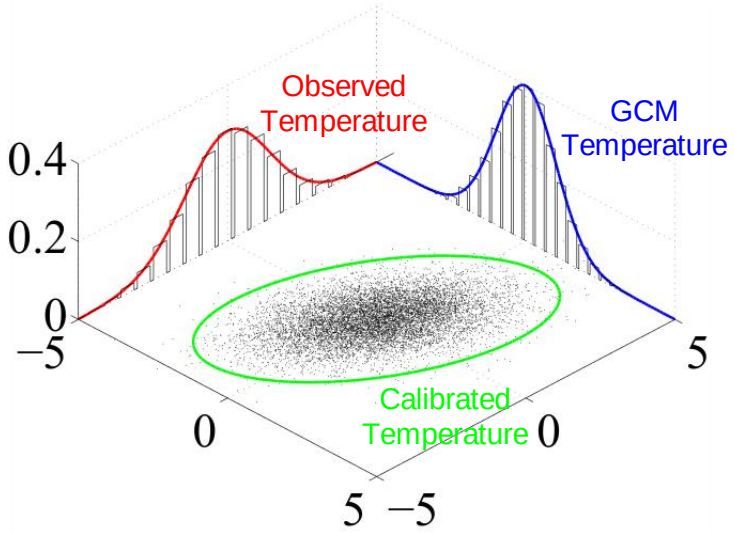
- **Calibration:** Relates dynamical model output to observed climate variables in order to correct both model bias and ensemble spread. Forecasts are returned to climatology in the absence of correlation between forecasts and observations.

- **BJP Calibration:** Statistically model relating the predictor (GCM output) and predictand (observations) using a continuous bivariate normal distribution
  - Use Monte-Carlo Markov-chain resampling technique to obtain 1000 ensemble members
  - Stated differently, we end up with 1,000 estimates of the relationship between observation and hindcast data and then generate a statistical ensemble of 1,000 forecasts

Overview: BJP Calibration

Bayesian Joint Probability (BJP) used in **Calibration, Bridging, and Merging (CBaM)** forecast system (Schepen et al. 2016; Strazzo et al. 2019) which provides NMME forecasts of temperature and precip over North America
https://www.cpc.ncep.noaa.gov/products/people/sstrazzo/cbam/index.php

- BJP models are developed using bivariate normal distributions to describe the relationship between a predictor and a predictand (i.e. GCM hindcasts and observations)
- Unlike other calibration methods, the parameters relating observed and hindcast data (e.g., means, covariances) are not viewed as fixed values. Instead, we use sampling methods to obtain a large sample (n=1,000) of possible parameters
- Stated differently, we end up with 1,000 estimates of the relationship between observed and hindcast data, which we can then use to generate a statistical ensemble of 1,000 forecasts



Yes this is from wikipedia but it's actually a nice visual!

Overview: Seasonal Calibration with BJP

Red shading indicates BJP calibration skill > raw for NMME (metric: BSS)

Project began by calibrating (as well as bridging and merging, not presented) dynamical model output from the North American Multi-Model Ensemble (NMME)

Figure shows the seasonal mean NMME Brier Skill Score (BSS) *difference* of BJP calibrated minus raw **2-meter temperature**

Figure courtesy Strazzo et al. 2019

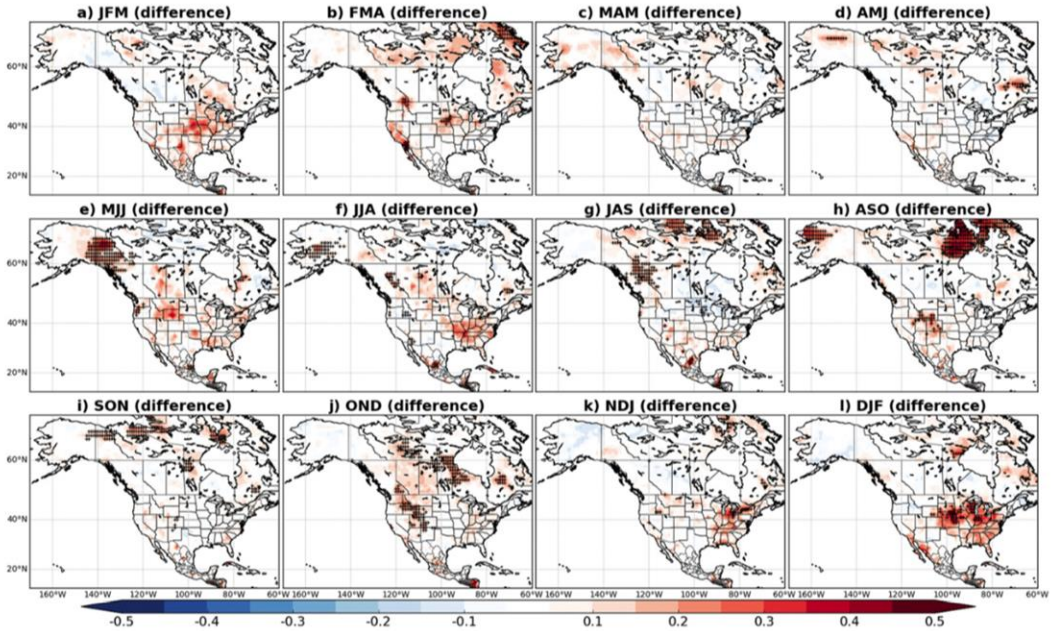


FIG. 3. Shading indicates Brier skill score differences between 1-month lead calibrated and 1-month lead raw forecasts of below-normal 2-m temperature for the NMME for each of the 12 overlapping 3-month seasons. Red shading indicates that calibrated forecast mean Brier skill scores exceeded raw forecast mean Brier skill scores over the 1982–2010 hindcast period. Hatching indicates significance at the 95% confidence level, as determined via a Wilcoxon rank sum test without accounting for field significance.

Example: Seasonal skill increase of BJP calibrated NMME lower tercile temperature compared to raw NMME (based on BSS)

Overview: Seasonal Calibration with BJP

Red shading indicates BJP calibration skill > raw for NMME (metric: BSS)

Project began by calibrating (as well as bridging and merging, not presented) dynamical model output from the North American Multi-Model Ensemble (NMME)

Figure shows the seasonal mean NMME Brier Skill Score (BSS) *difference* of BJP calibrated minus raw **precipitation**

Figure courtesy Strazzo et al. 2019

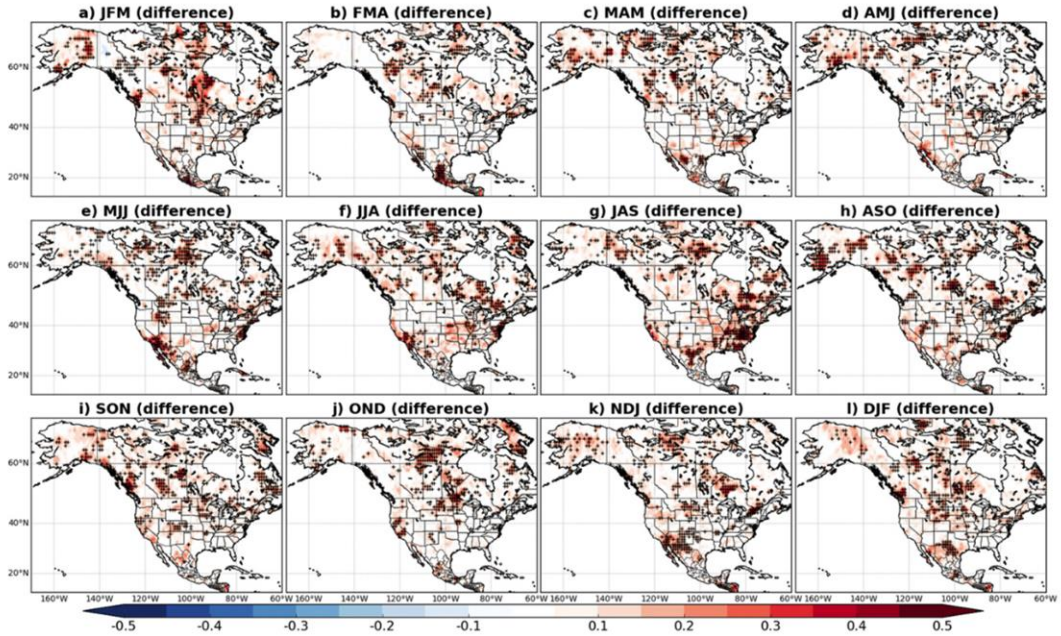
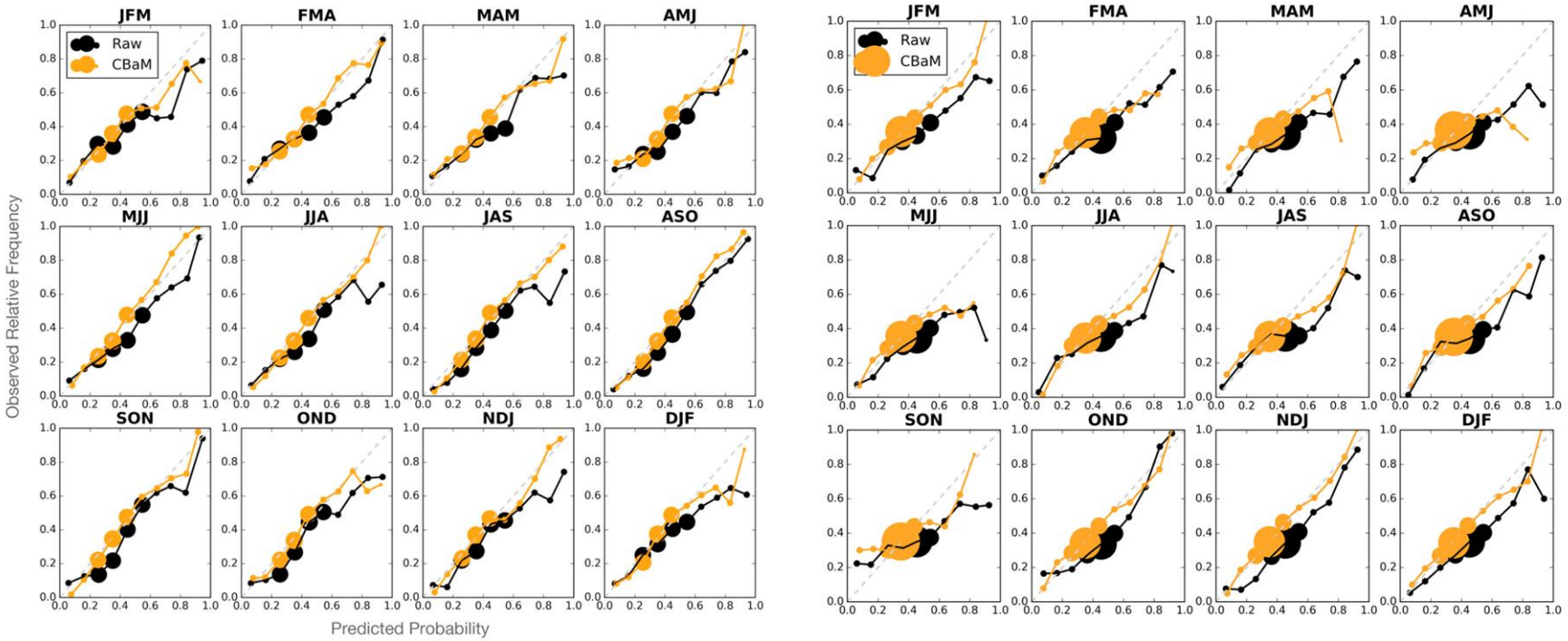


FIG. 10. Shading indicates Brier skill score differences between 1-month lead calibrated and 1-month lead raw forecasts of below-normal precipitation rate for the NMME for each of the 12 overlapping 3-month seasons. Red shading indicates that calibrated forecast mean Brier skill scores exceeded raw forecast mean Brier skill scores over the 1982–2010 hindcast period. Hatching indicates significance at the 95% confidence level, as determined via a Wilcoxon rank sum test without accounting for field significance.

Example: Seasonal skill increase of BJP calibrated NMME lower tercile precipitation compared to raw NMME (based on Brier Skill Score, BSS)

Reliability: Seasonal Calibration with BJP



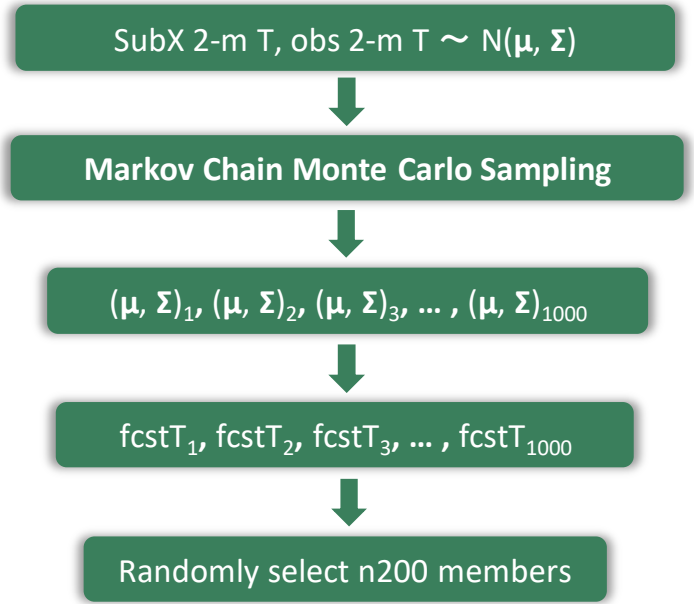
The benefits of calibration extend beyond skill metrics, and calibration also adds to reliability. Above shows the reliability of raw (black) and BJP calibrated (orange) NMME 2-meter temperature and precipitation (1 month lead)

Overview: Calibrating SubX with BJP

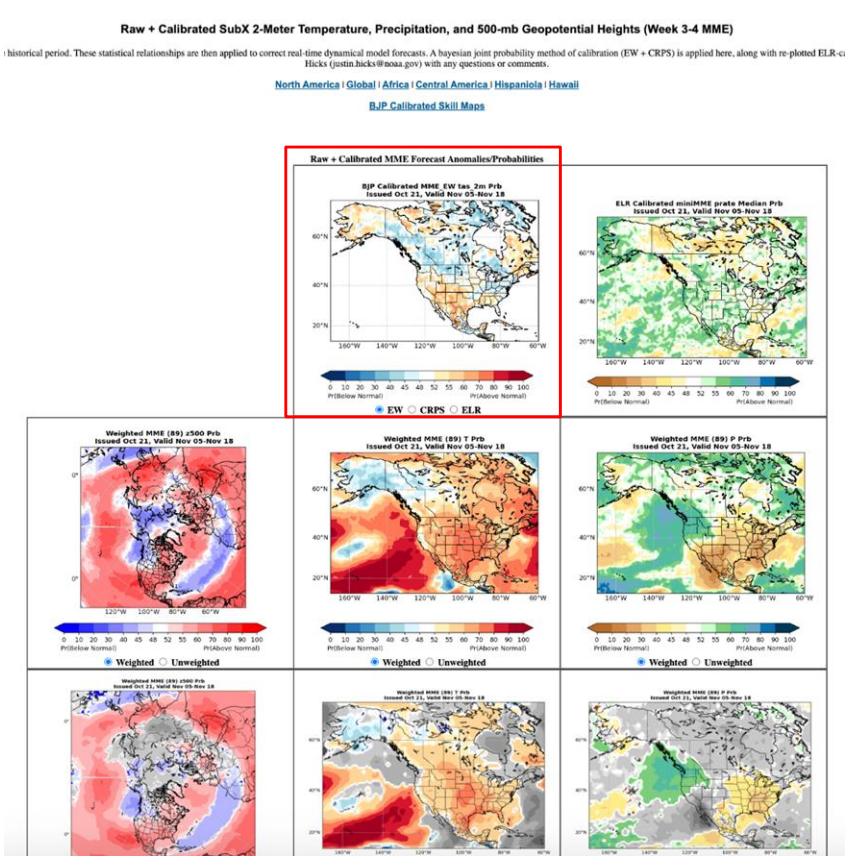
Given the success on the seasonal timescale, we applied BJP calibration to the subseasonal timescale

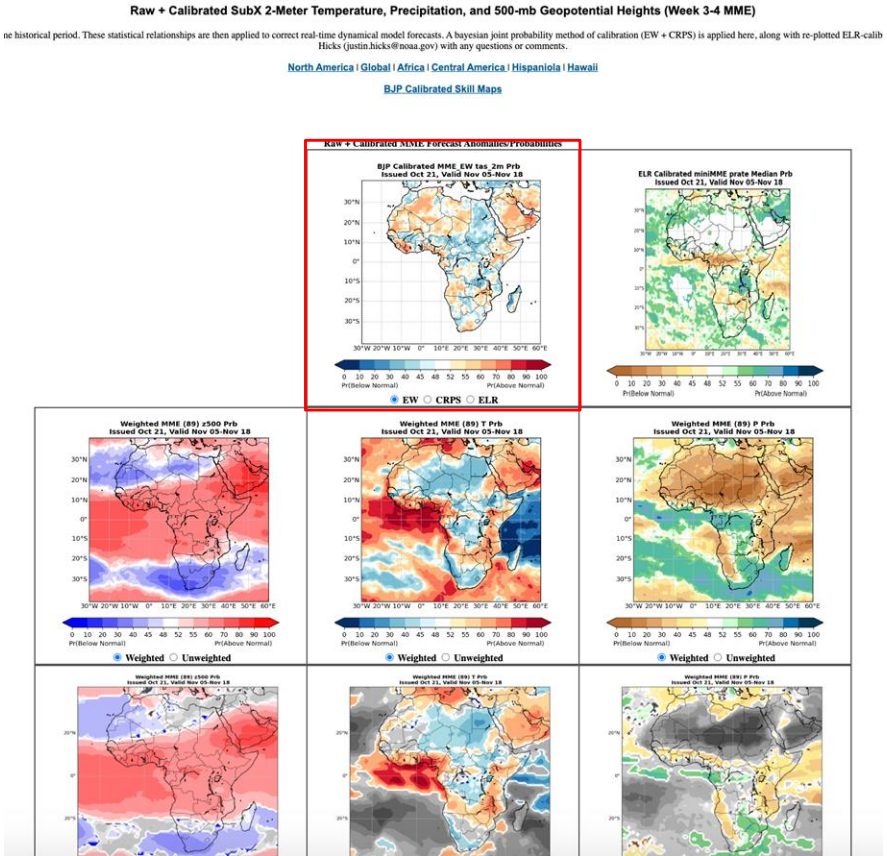
- BJP calibration applied to a 3-model miniMME (multi-model ensemble) from SubX (experimentally)
  - EMC-GEFSv12, ESRL-FIMv2, NCEP-CFSv2 (models initialized on Wednesdays)
- Calibration set up to align with operational CPC forecasts
  - Operational forecasts issued on Friday for days 15-28 (mean)
  - Given Wednesday initialization from SubX models, calibrate the mean 17-30 day forecasts
- Develop one calibration model for each week of the year, for each dynamical model (52*3 calibration models in total)
- Apply leave-one-year-out cross validation
- The calibrated output is combined into an MME by equally weighting the 3 models or by using weighting determined by the the Continuous Rank Probability Score (CRPS)
- Calibration performed for hindcasts and also in realtime for the 3 models & provided to forecasters on Fridays.
- Experimentally applied to two category 2-meter temperature *only* (tercile is currently in development)

Bayesian Joint Probability (BJP) Model

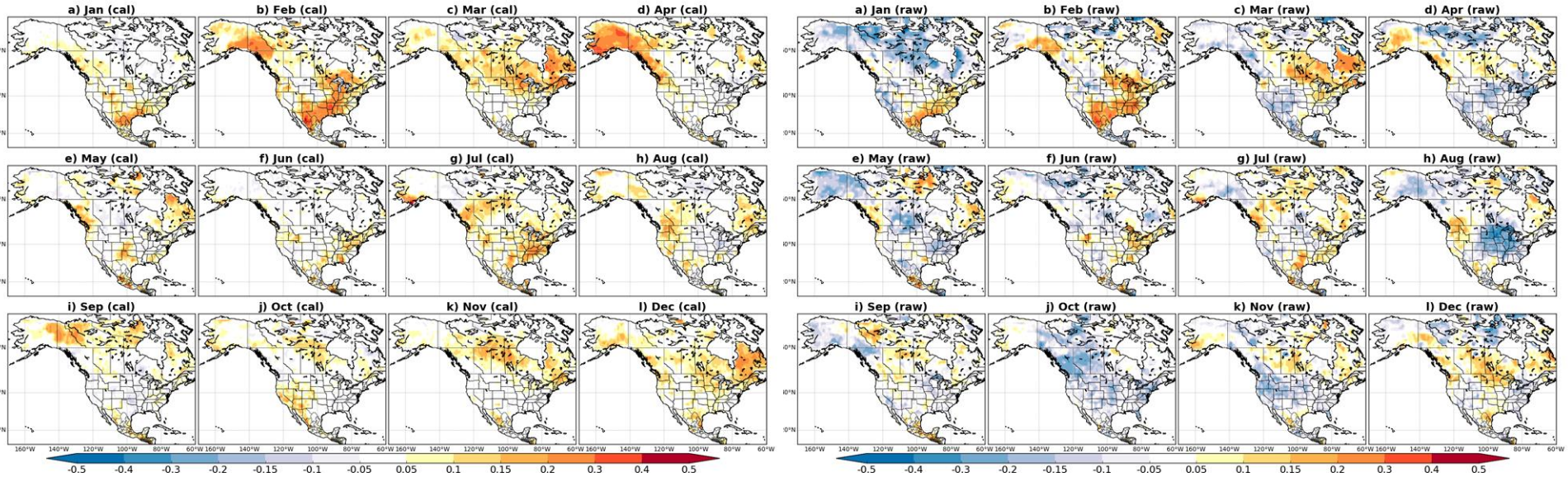


Examples of Calibrated and Raw SubX Forecasts



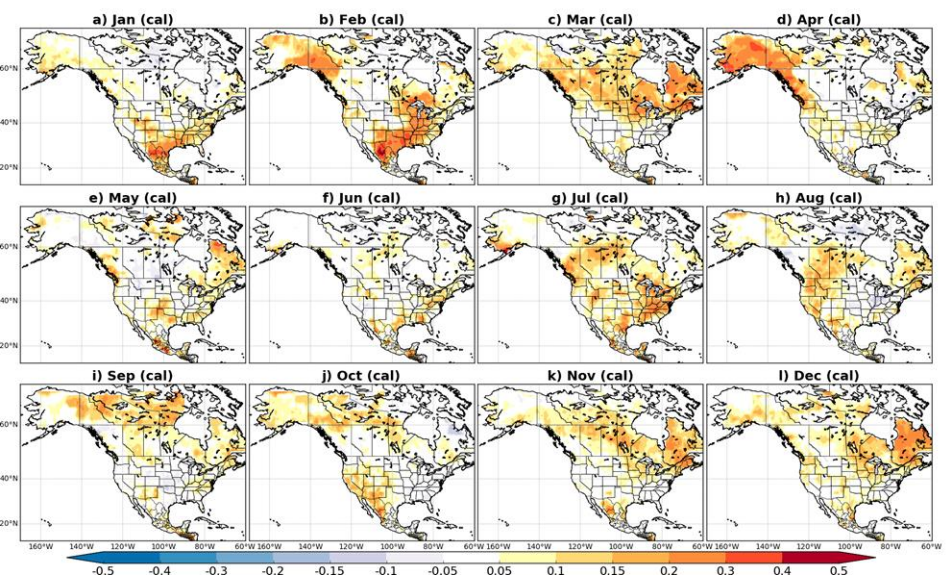


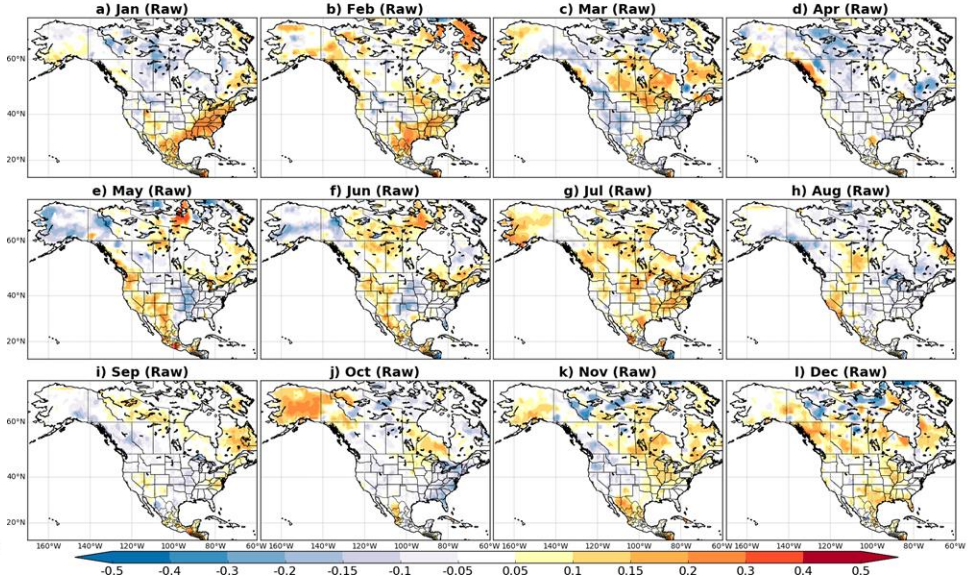
Subseasonal Skill: BJP Calibrated vs. Raw GEFSv12



BSS for 2-category BJP calibrated (left) and raw (right) GEFSv12 Week 3-4 2-meter temperature hindcasts
Monthly mean depicted for simplicity.

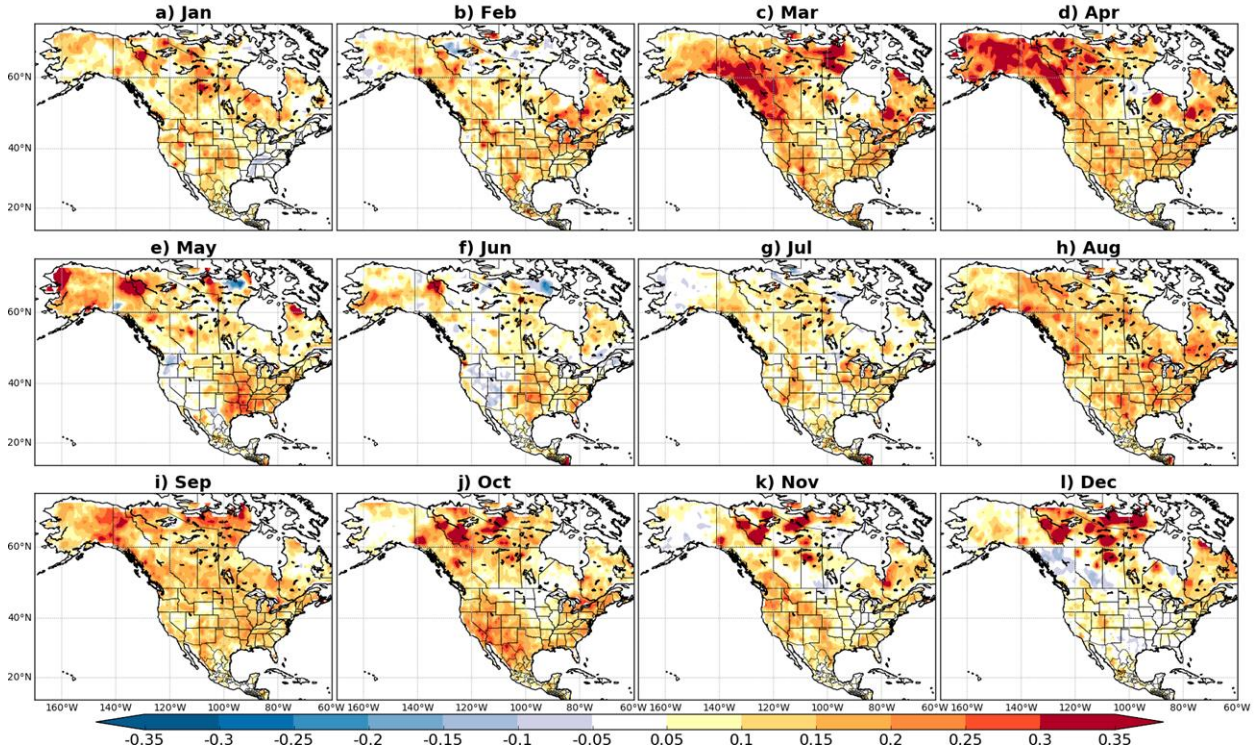
Subseasonal Hindcast Skill: Calibrated vs. Raw MME





BSS for 2-category BJP calibrated (left) and raw (right) miniMME Week 3-4 2-meter temperature hindcasts
Monthly mean depicted for simplicity.

Subseasonal Hindcast Skill: Calibrated minus Raw MME



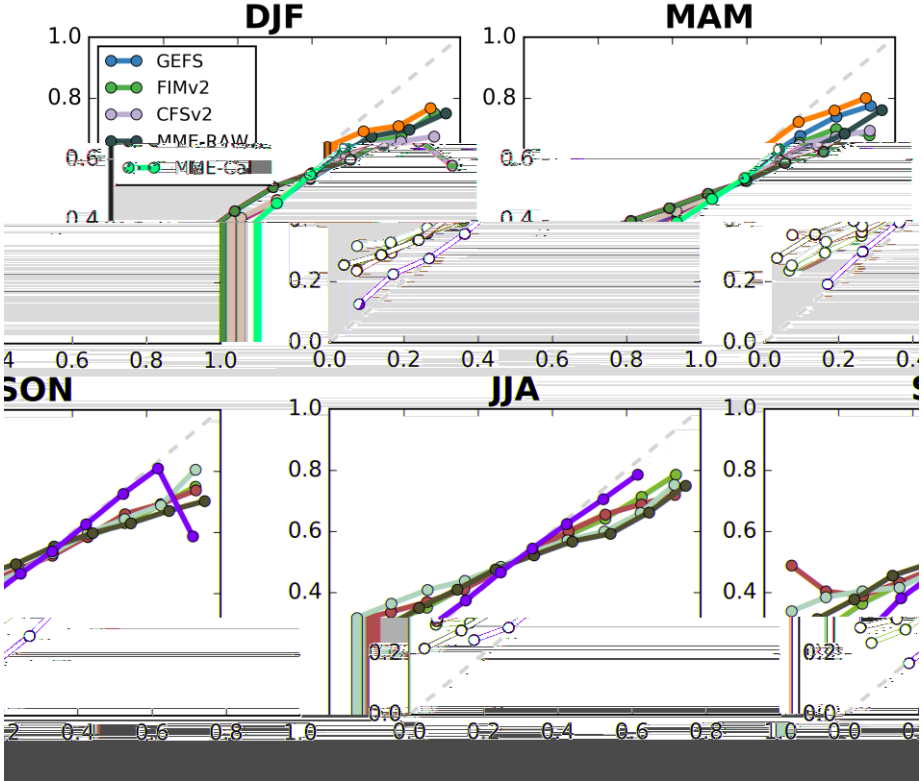
BJP calibrated minus raw miniMME

(red shading indicates BSS is higher for BJP calibrated and vice versa)

Reliability Enhancements due to BJP Calibration

Comparison of Reliability for 2-Category Raw and BJP Calibrated mini-MME tmp2m Forecasts

- **Calibrated MME more reliable** than calibrated **SubXGEFS**, **FIMv2** or **CFSv2**, (small ensemble size), or **MME member count (raw)** probability in all seasons

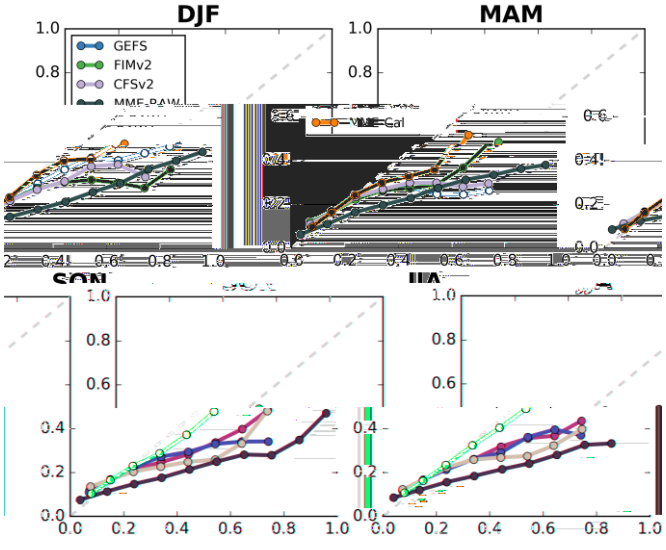


A more reliable forecast will lie closer to the diagonal, i.e. forecasts with a probability of X% are consistent with the observed outcome occurring X% of the time

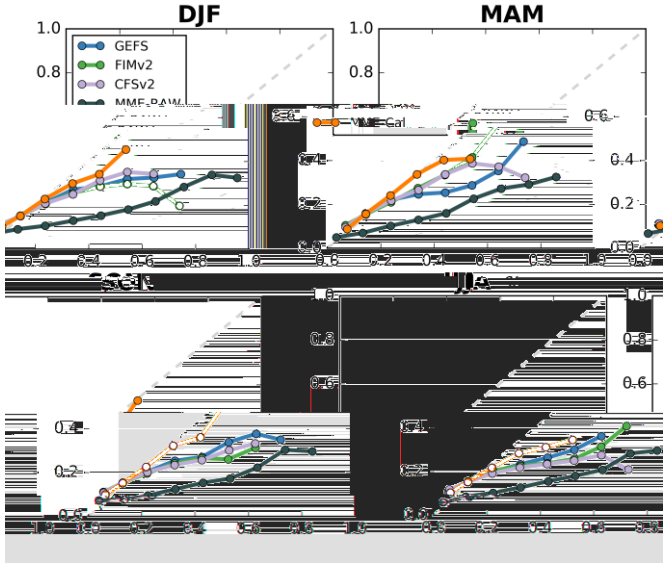
Reliability Enhancements due to BJP Calibration

Extreme above/below normal reliability (15th / 85th percentiles)

- **Calibrated MME** essential to reliability of probabilities of extremes
- **Raw MME** has much less reliable probabilities
- Individual calibrated **SubXGEFS**, **FIMv2** or **CFSv2** are less reliable than **MME**



Extreme **below** normal



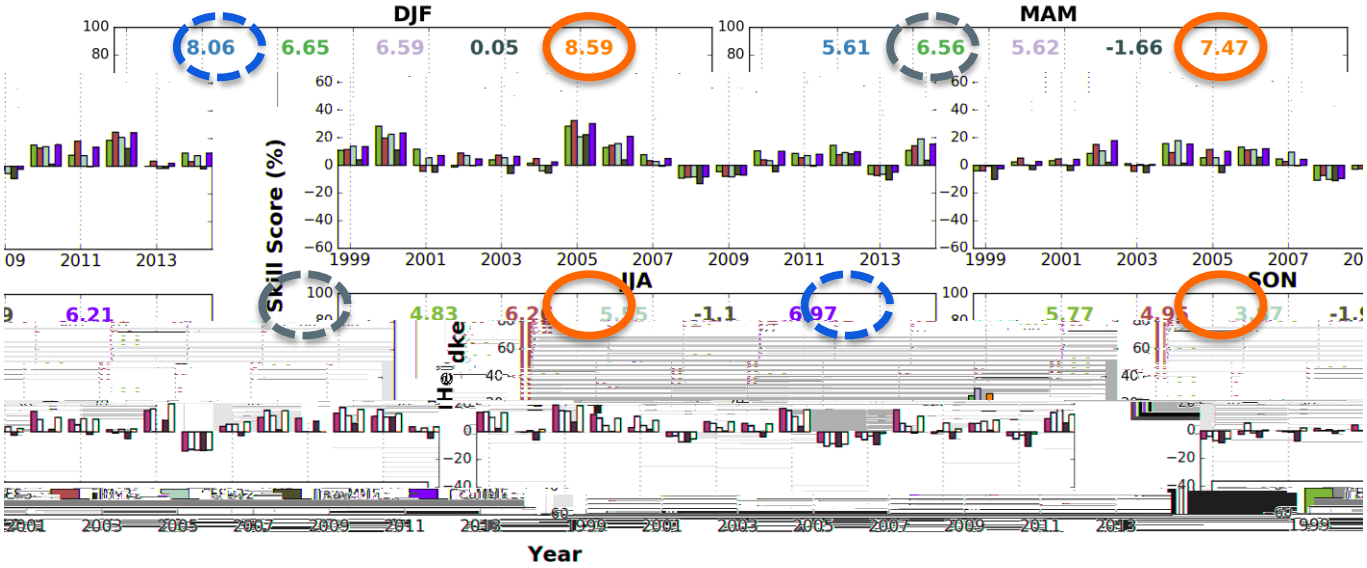
Extreme **above** normal

Heidke Skill Score Increases due to BJP Calibration

Extreme above normal reliability (85th percentile)

1st & 2nd ranked models

- **Calibration** of **raw mini-MME probabilities** *improves overall Heidke Skill Score*
- **Raw mini-MME** has less reliable probabilities AND lower hit rate
- MME more skillful in most months / years than **GEFS**, **FIMv2** or **CFSv2**



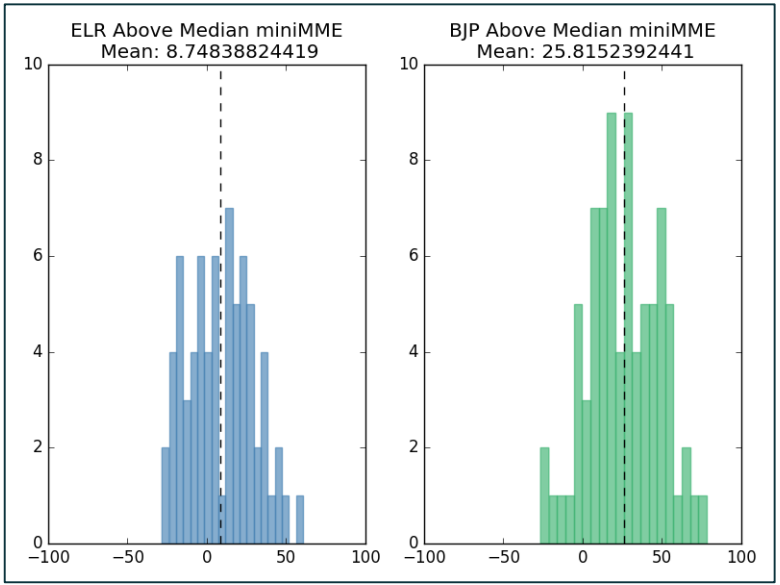
Courtesy of S. Strazzo

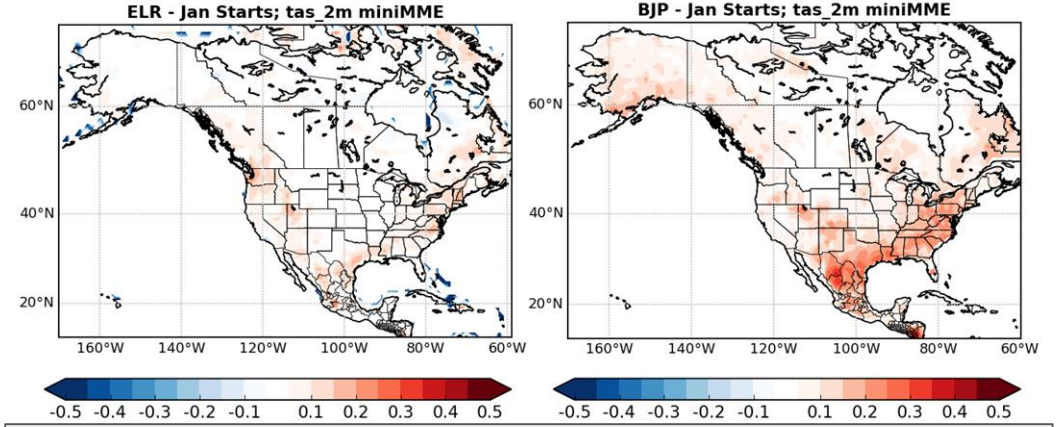
Other SubX Calibration Methods: Extended Logistic Regression (ELR)

A joint Climate Test Bed (CTB) project between the International Research Institute (IRI) and Climate Prediction Center (CPC)

Extended Logistic Regression forecasts run by IRI and provided to CPC
http://iridl.ldeo.columbia.edu/maproom/Global/ForecastsS2S/index.html

Comparison of BJP vs. ELR Heidke Skill Scores (HSS) and Brier Skill Scores (BSS)
The key difference between the methods is use of Bayesian vs. regression techniques





BJP outperforms ELR calibration, though both serve to improve skill and calibration training observations differed, possibly leading to some differences in verification

BJP vs. Probability Anomaly Correlation (PAC) Calibration Method (Seasonal)

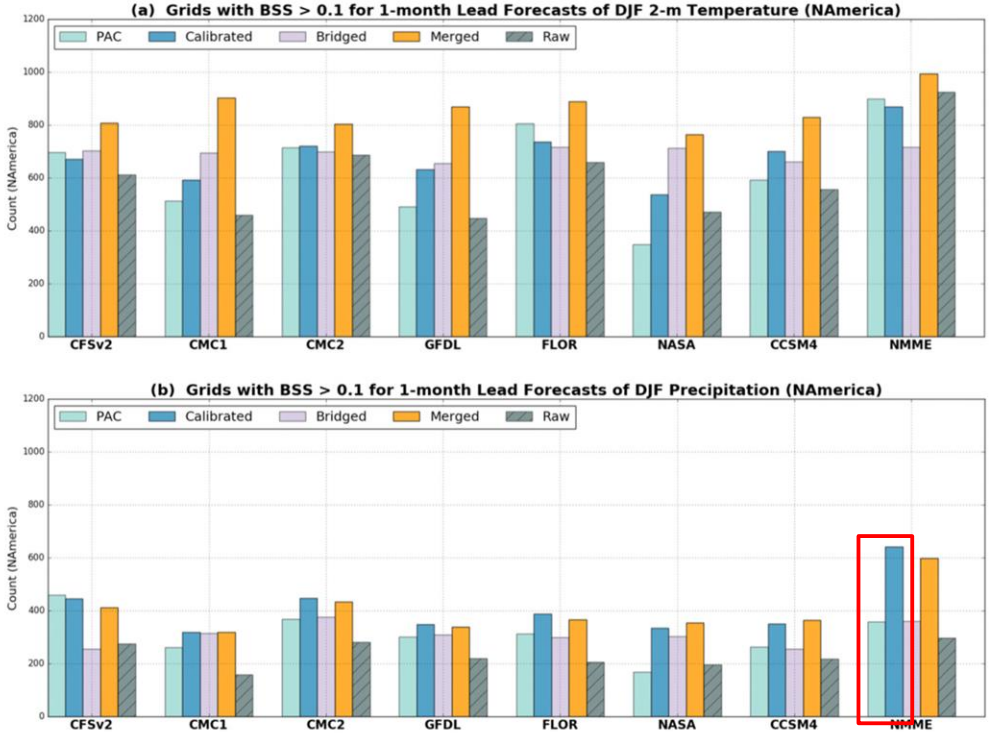


FIG. 7. The number of grid cells over North America with BSS > 0.1 for PAC-calibrated (light blue), BJP-calibrated (dark blue), bridged (purple), merged (orange), and raw (gray) 1-month lead forecasts of below-normal DJF (a) temperature and (b) precipitation from each of the NMME member models and the multimodel mean.

Though this result is for seasonal 2-meter temperature and precipitation, BJP calibration tends to meet or outperform PAC calibration, particularly for precipitation.

2-meter temperature results vary by model.

Upcoming Goals

- Currently, BJP calibration is *experimental* and applied to both experimental models and operational models
- Given the clear gains in skill of BJP calibrated 2-meter temperature over raw model output on the subseasonal timescale, as well as BJP calibrated precipitation on the seasonal timescale, our goals are to:
  - Apply the BJP methodology to operational models that support week 3-4 forecasts at CPC; including CFSv2, ECMWF, GEFSv12, and, when available, the Unified Forecast System (UFS).
    - Note that current datastreams for BJP-SubX include CFSv2 and GEFSv12, but these are from non-operational datastreams. We will transition to operational datastreams for these models.
  - Apply BJP methodology that has been extensively tested on seasonal precipitation to week 3-4 precipitation from CFSv2, ECMWF, GEFSv12, and UFS.
  - Transfer BJP code and modules to the CPC Compute Farm (CF).
- Our current project through the NOAA Internal Research to Operations call from JTTI is expected to meet these goals.