Temporal disaggregation of seasonal temperature forecasts from Bayesian Joint Probability (BJP) calibrated NMME to predict daily extremes

Johnna Infanti, Dan Collins, Sarah Strazzo, Andrew Schepen, QJ Wang
NOAA/CPC, Innovim, Embry-Riddle Aeronautical University, CSIRO, University of Melbourne

Introduction

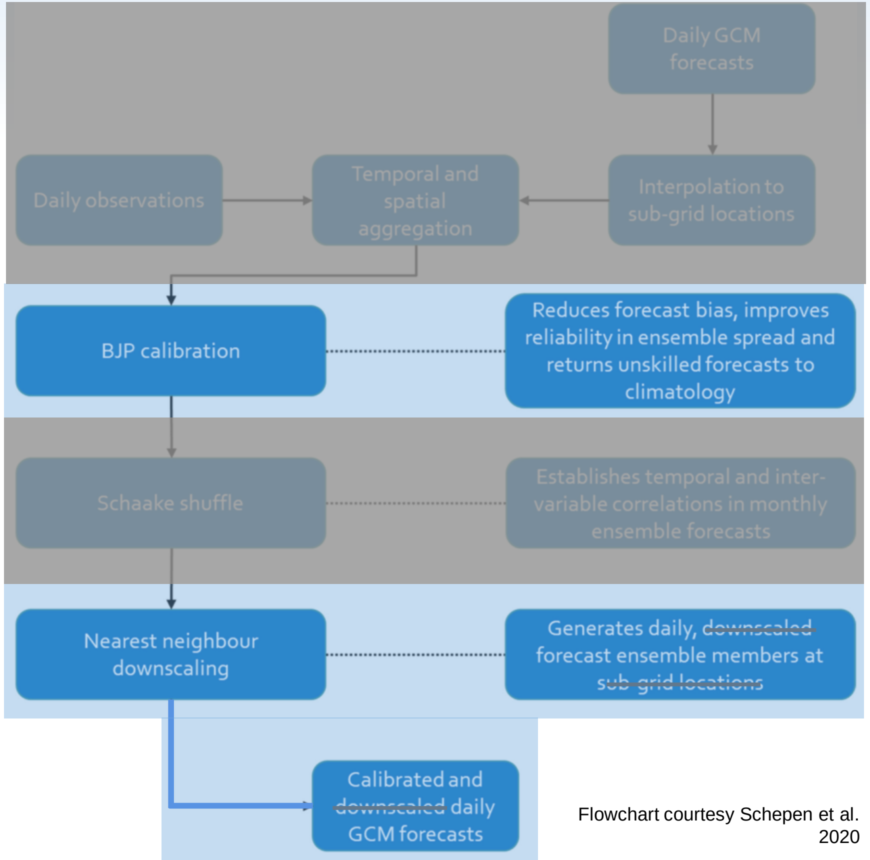
While seasonal forecasts have high value for many sectors and stakeholders, they are typically issued at coarse temporal (and spatial! Though here we focus on temporal) scales

Also, GCM hindcasts/forecasts can be biased, so calibration methods are sometimes needed or desired to adjust raw forecasts to minimize this bias, improve reliability, and improve skill

We utilize (modified) forecast-calibration multivariate-downscaling (FCMD) (Schepen et al. 2020) to ***temporally disaggregate*** (i.e. separate or break apart) raw and calibrated seasonal forecasts to daily

Overarching Goal: To provide forecasts of the distribution of daily values within a given season, that preserve the statistical properties awarded by calibration and historical daily sequences. Forecast probability of extreme days (PoEx) within the season.

Methods: Overview



Flowchart courtesy Schepen et al. 2020

Schepen et al. 2020 disaggregated and downscaled seasonal gridded forecasts to daily/stations over Australia using the FCMD technique which has its basis in the Method of Fragments (MoF) methodology.

Data we are disaggregating

Raw CFSv2 hindcasts of 2-meter temperature over North America (24 ensemble members)

Bayesian Joint Probability (BJP) Calibrated CFSv2 hindcasts of 2-meter temperature over North America (100 ensemble members)

Calibrated with Global Historical Climatology Network and Climate Anomaly Monitoring System (GHCN-CAMS)

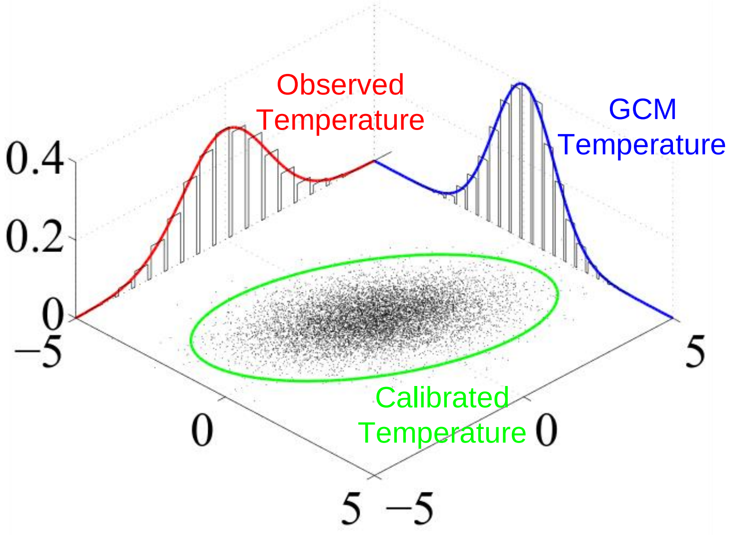
Raw and calibrated data are disaggregated with respect to the CPC daily global temperature dataset (GLBT)

Bayesian Joint Probability (BJP) Calibration Crash Course

As noted earlier, we are disaggregating raw and **calibrated** seasonal hindcasts to daily… so what is the calibration method used?

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

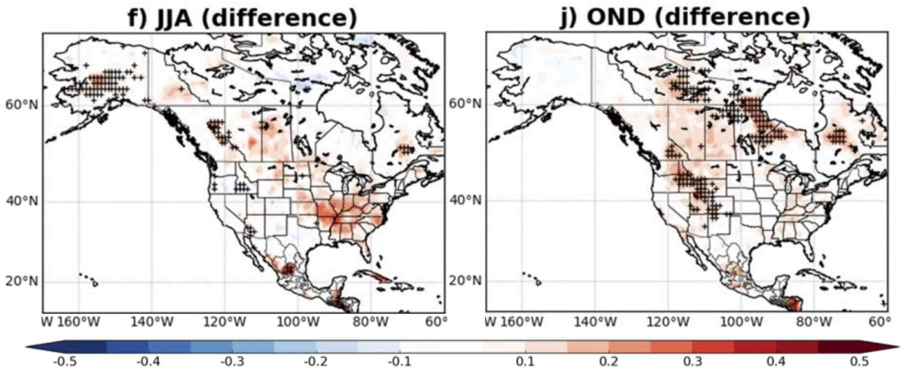
- Calibration models, in general, are developed using observed and hindcasts data
- BJP models are developed using bivariate normal distributions to describe the relationship between a predictor and a predictand (i.e. GCM hindcasts and observations)
- Unlike other calibration methods, the parameters 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!

Bayesian Joint Probability (BJP) Calibration Crash Course

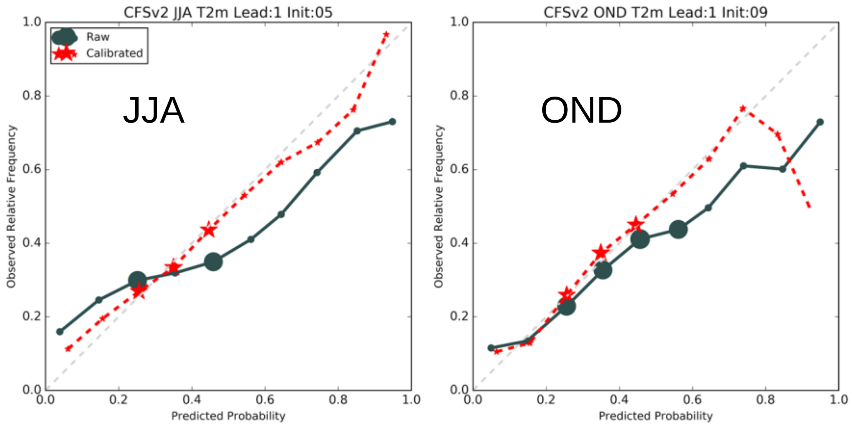
How does calibration help with skill?



Calibrated < Raw

Calibrated > Raw

Brier skill score (BSS) differences between 1-month lead calibrated and 1-month lead raw forecasts of below-normal 2-m temperature for the NMME
Figure courtesy Strazzo et al. 2019



Upper tercile 2-meter temperature raw and calibrated reliability for CFSv2 (lead 1)

Methods: CFSv2 Disaggregation

Schepen et al. (2020) apply the **method of fragments technique (MOF)** to downscale rainfall, tmin, tmax, and solar radiation simultaneously. With calibration, the authors call this entire technique **forecast calibration-multivariate downscaling (FCMD).** *Again, we focus on 2-meter temperature and temporal disaggregation.*

CFSv2 Raw or Calibrated Seasonal 2-meter Temperature Forecast for a given season/year

24 ensemble members for Raw
100 Ensemble members for BJP
Run through entire hindcast period of 1982-2010

Standardize forecast temperature based on observed mean and standard deviation

Find all the dates when the squared "error" between observations and forecasts is smallest (analog dates)

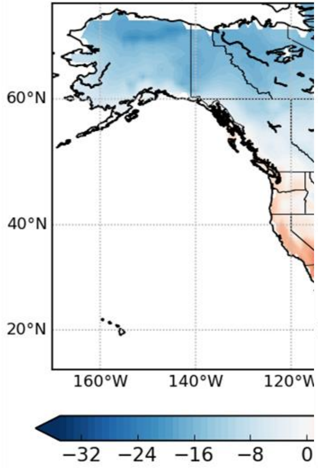
For each analog date, calculate the weight of that day in the given season; i.e. observed daily temp from analog date search/observed seasonal mean) - This gives you weights for n # of days in a season e.g. 92 days in June-Aug, etc.

Multiply the CFSv2 seasonal forecast by each of these weights to form your disaggregated daily forecast
Gives e.g. 92 days x 100 ensemble members
Examples shown for **Oct-Dec** and **Jun-Aug**

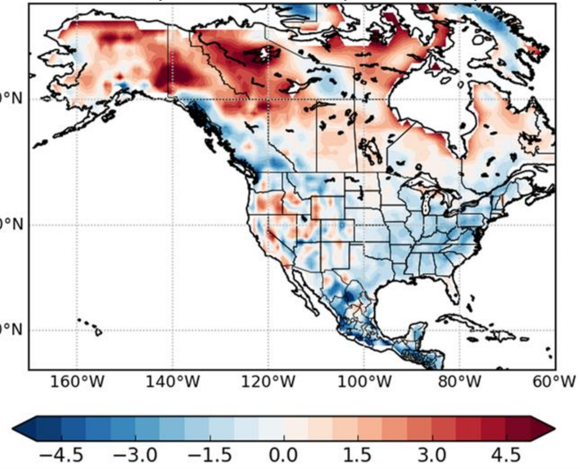
Benefits of MoF for Disaggregation

- Seasonal skill and statistics are maintained when disaggregating, including any benefits gained from calibration
- Thus, we will have the distribution of days within a season that matches the statistics of the forecasted season (rather than, for example, using a different model with daily data, etc)
- Multivariate and can correct for the relationship between variables (e.g. temperature and precipitation; examples here for temperature, future work will involve multivariate)

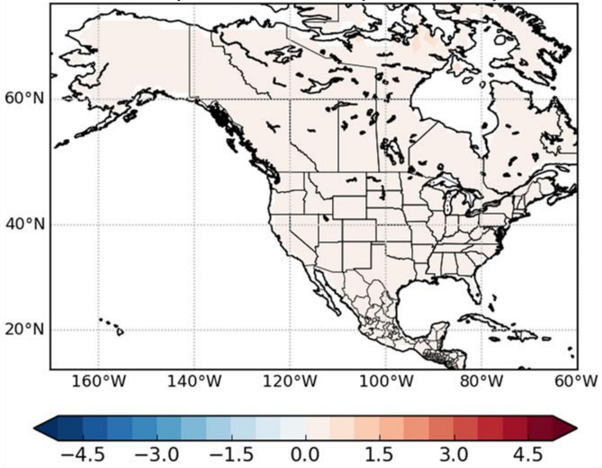
Observed Oc
Temperature Cli



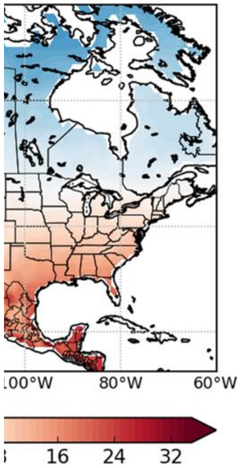
Raw - Obs Difference
Temperature Climo (1982-2010)



BJP - Obs Difference
Temperature Climo (1982-2010)

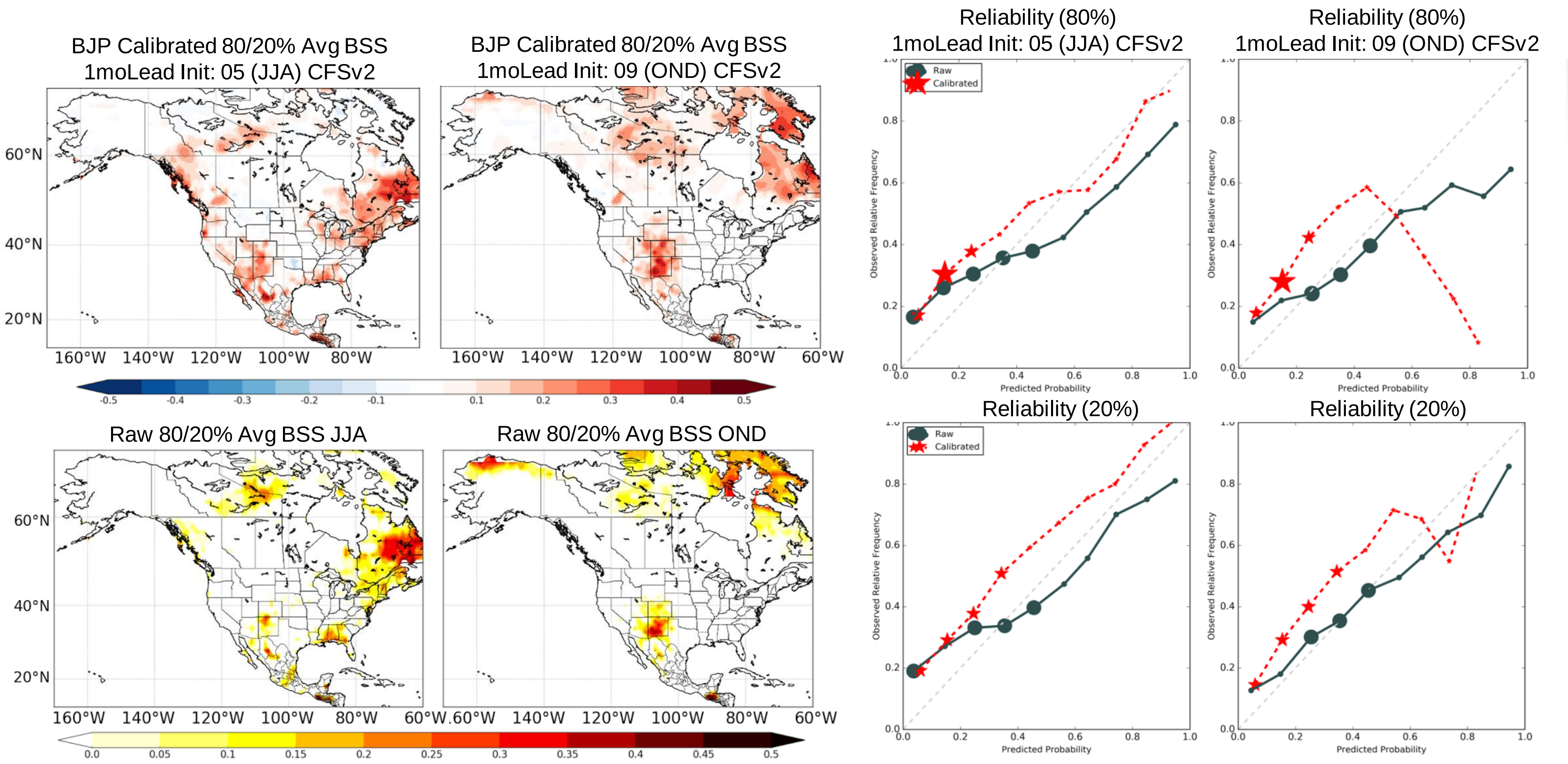


d OND
(1982-2010)



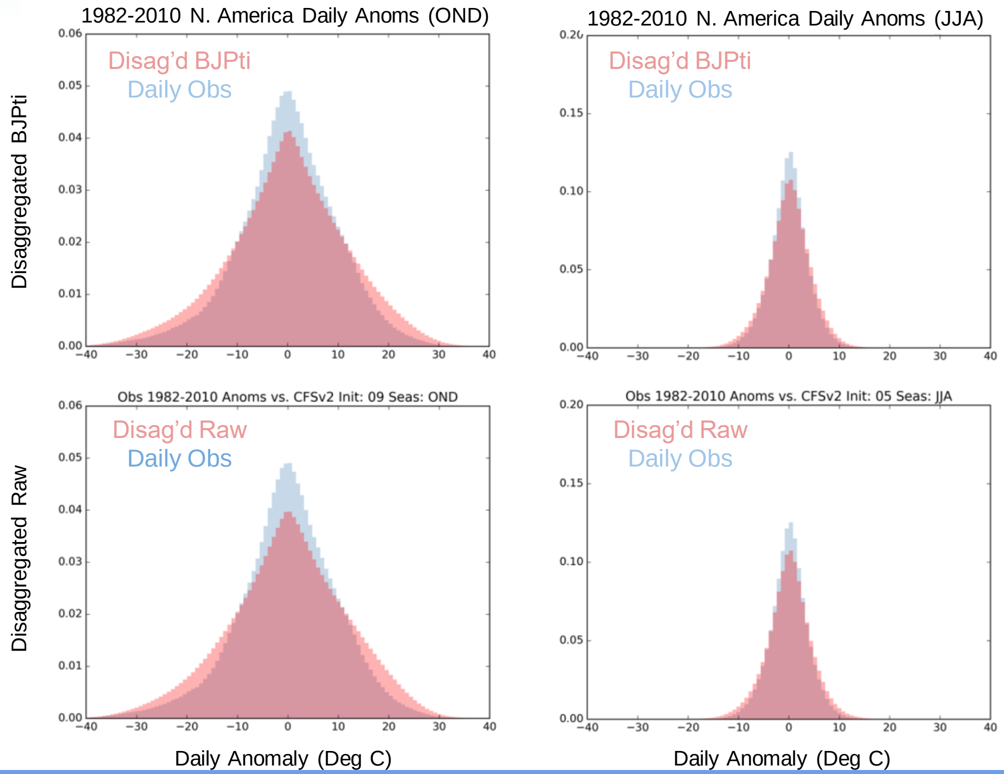
Results for CFSv2

Skill of seasonal “extremes” (80% and 20%)



Daily distribution within the given season

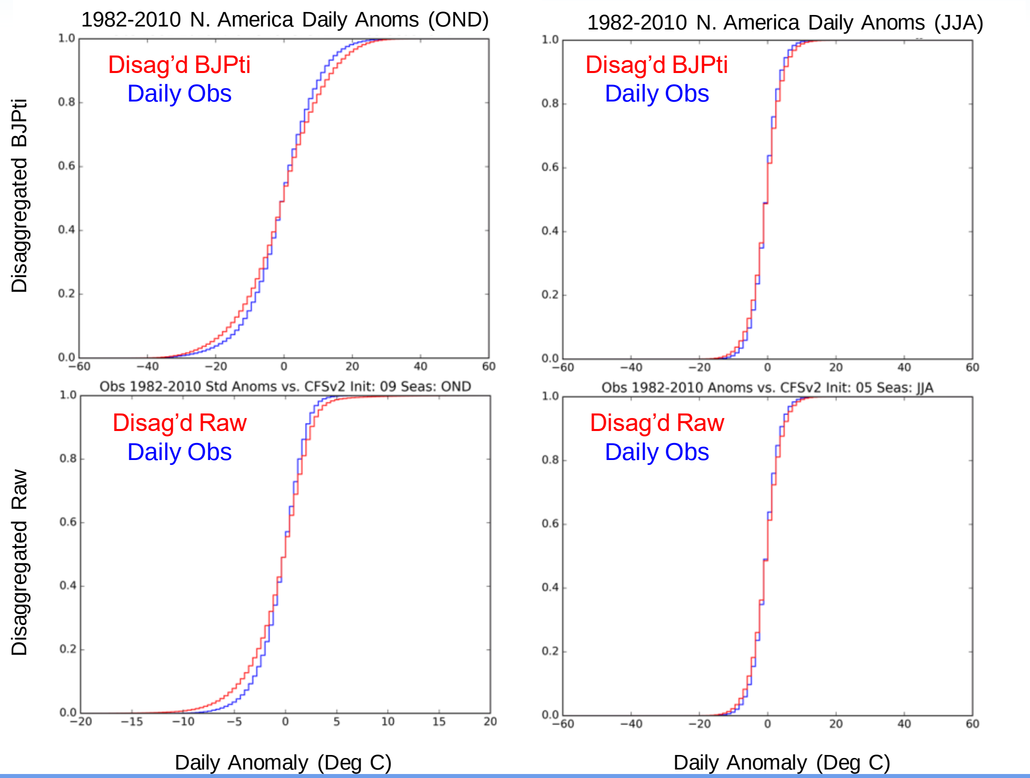
Similarly to Schepen et al. 2019, we want to determine if the distribution of days within a season matches observations once we disaggregate - Especially the extremes (tails)!



- While the distribution isn't "bad", we see similar issues in both calibrated and raw disaggregated data, e.g.
  - Lower neutral events
  - Higher extremes during winter (could see lack of reliability)
  - Summer looks slightly better than winter, but the skill was also slightly better
- Note that this is for the entirety of North America, regional differences, particularly where there is skillfulness, may be different

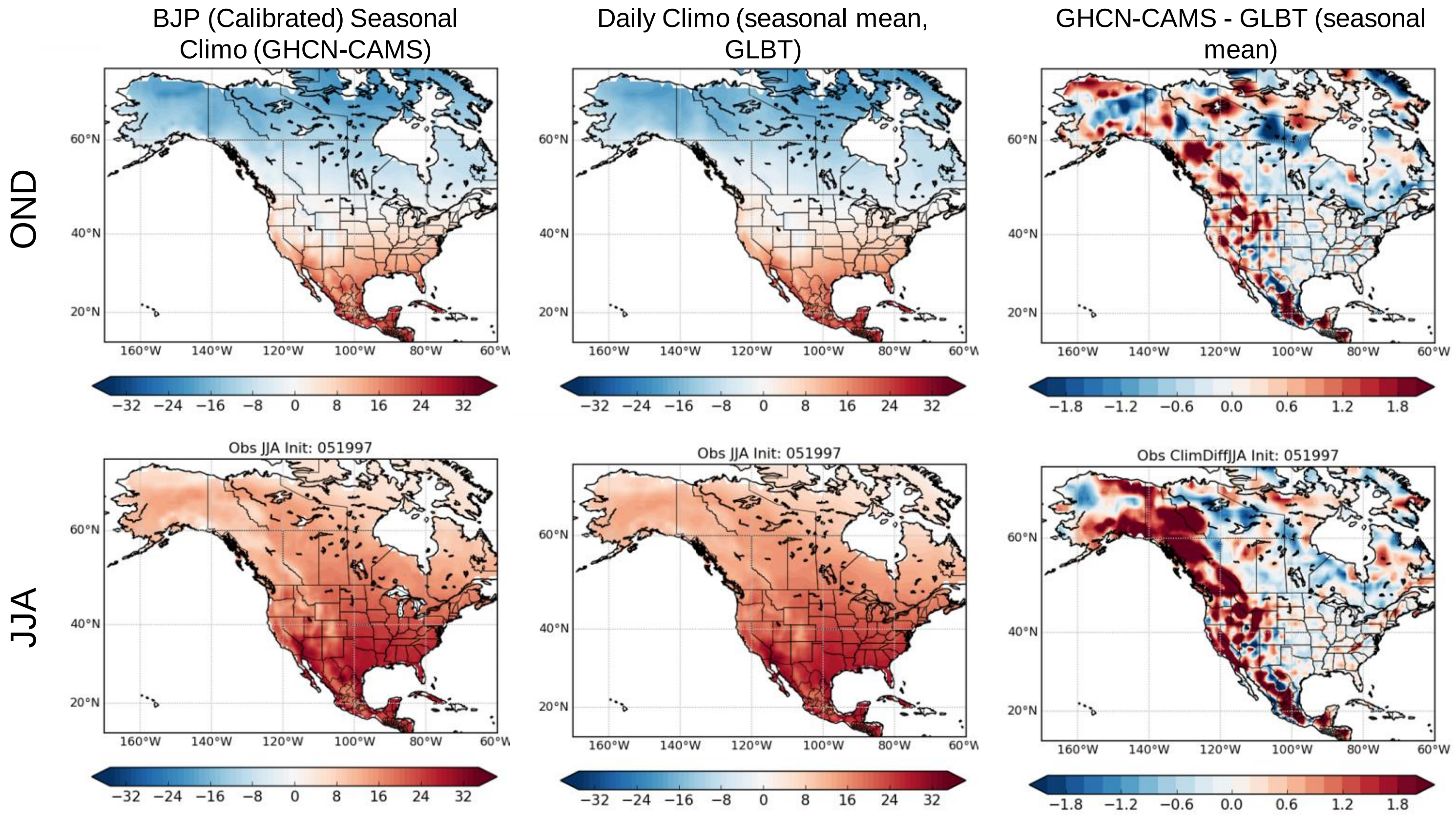
Daily distribution within the given season

Similarly to Schepen et al. 2019, we want to determine if the distribution of days within a season matches observations once we disaggregate - Especially the extremes (tails)!



CDF's show a similar result but it might be slightly easier to see the extremes

Difference in Calibrated vs. Daily Climatologies



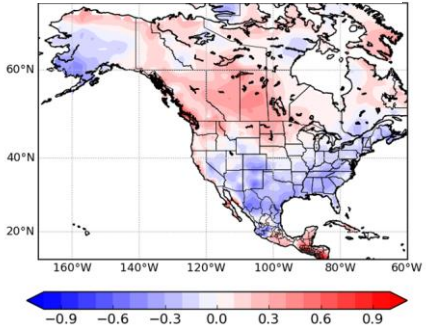
Example of a forecast and what we can do with these data...

Reminder!

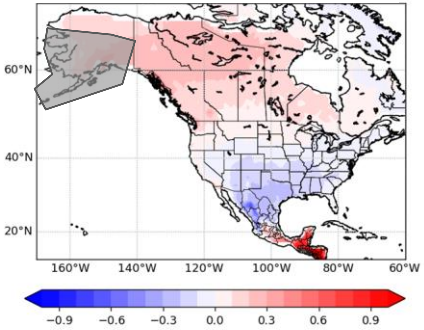
Overarching Goal: To provide calibrated forecasts of the distribution of daily values within a given season, that preserve the statistical properties awarded by calibration and historical daily sequences. Forecast probability of extreme days (PoEx) within the season.

Example Forecast: OND1997 (while there is lower skill in the winter months, it is likely that this was a forecast of opportunity, so used here for demonstration purposes)

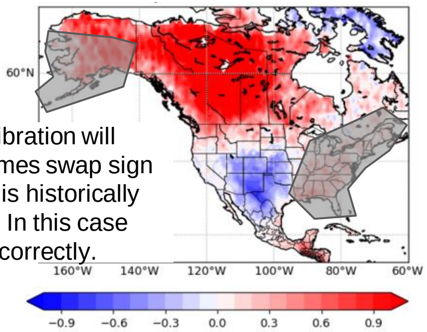
Observed Anomaly OND1997



CFSv2 Raw Anom OND1997



CFSv2 BJP Anom OND1997

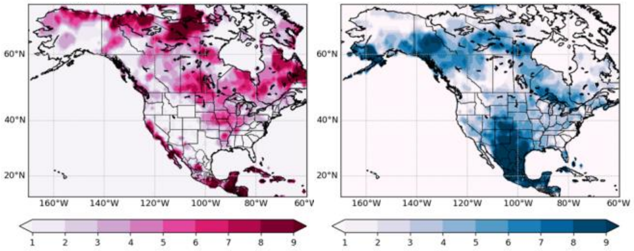


Calibration will sometimes swap sign if skill is historically low - In this case incorrectly.

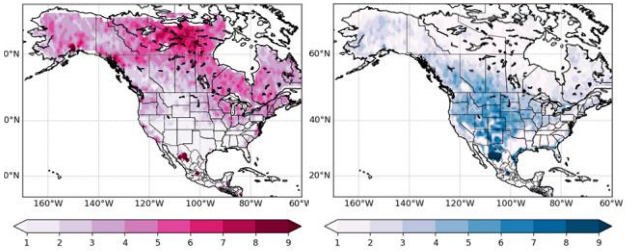
Disag 24 ensemble members

Disag 100 ensemble members

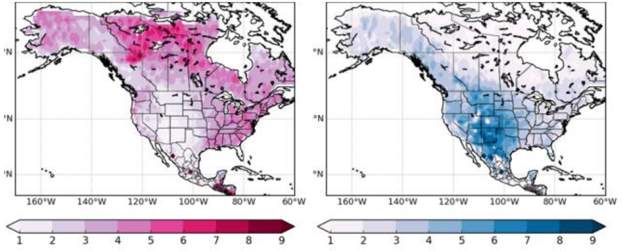
Obs # of days GT or LT 95th percentile threshold



Raw Ens. Median # of days GT or LT 95th percentile threshold



BJP Ens. Median # of days GT or LT 95th percentile threshold



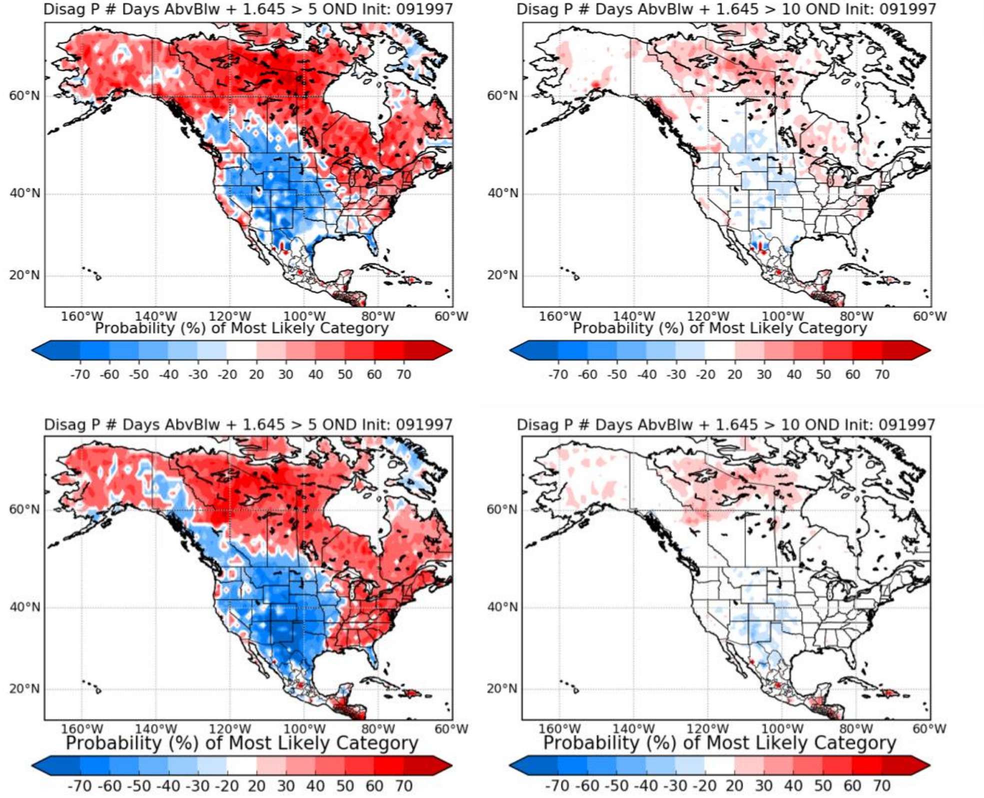
Example of a forecast and what we can do with these data...

PoEx (95th percentile)
5 or more days within the season

PoEx (95th percentile)
10 or more days within the season

Raw OND1997 Disag

BJP OND1997 Disag

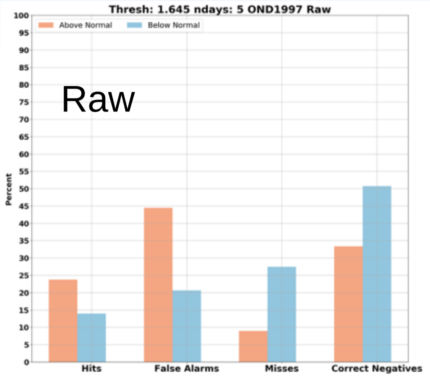


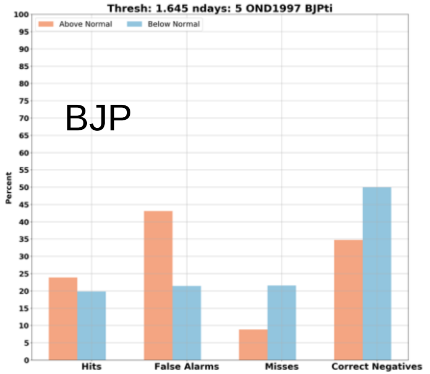
How to Read:
Probability of 5 (10) or more "extreme" days in the given season
Blue: Probability is higher for lower extreme
Red: Probability is higher for higher extreme

- This is just an example forecast
- Thresholds can be shifted or changed based on user/forecaster needs easily
- Can achieve things like:
    - # of days within season > a given temperature
    - # of consecutive days of extreme heat or cold
    - Regionally defined thresholds
    - & more!

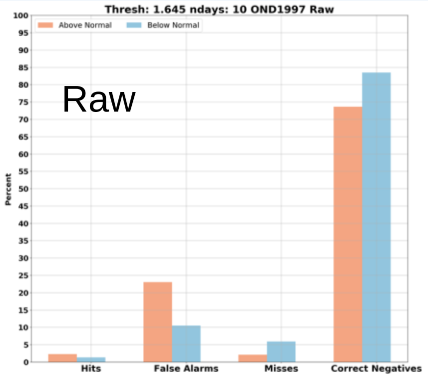
Verification of Example Forecast

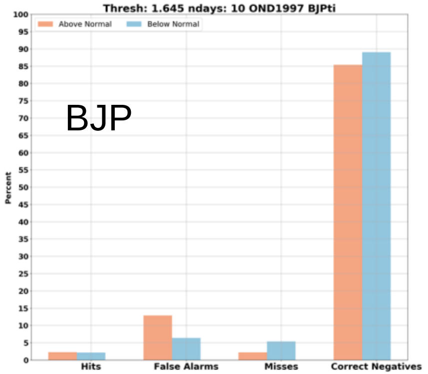
Contingency Table for
PoEx > 5 days





Contingency Table for
PoEx > 10 days





	> 5 Days in Top Extreme		> 10 Days in Top Extreme	
	Raw	**BJP**	**Raw**	**BJP**
Hit	23.76%	23.88%	2.25%	2.29%
False Alarm	44.47%	43.10%	23.04%	12.91%
Miss	8.93%	8.85%	2.09%	2.21%
Correct Negative	33.33%	37.74%	73.66%	85.40%
Threat Score HR/(HR+Miss+FA)	30.79%	31.49%	8.22%	13.15%

	> 5 Days in Bot Extreme		> 10 Days in Bot Extreme	
	Raw	**BJP**	**Raw**	**BJP**
Hit	13.95%	19.83%	1.29%	2.17%
False Alarm	20.67%	21.43%	10.49%	6.39%
Miss	27.42%	21.55%	5.91%	5.35%
Correct Negative	50.74%	49.98%	83.51%	89.06%
Threat Score HR/(HR+Miss+FA)	22.49%	31.57%	7.29%	15.60%

Green squares indicate the better score. Results are mixed for lower extreme/5 days, but BJP is the overall winner. Note that the # of gridpoints with an extreme is low for >10 days, so we want the Correct Negs to be high! Threat score is a measure of accuracy, how will the forecast “yes” events correspond to observed “yes” events?

Concluding Remarks
(and what we hope is in store for the future)

Concluding Remarks

- We have applied the methodology from Schepen et al. 2019 (used for ECMWF forecasts over stations in Australia) to North American CFSv2 raw and calibrated temperature forecasts to statistically disaggregate seasonal forecasts to daily
- Goal of disaggregation is to provide a forecast of the distribution of days within a season and probability of extreme days (PoEx), that matches and preserves the statistics of the season
- Overall, the disaggregated raw and calibrated 1982-2010 hindcasts showed a respectable comparison to observed distribution, with a few exceptions, and noting that the calibrated forecasts are calibrated to a different observed dataset than used for verification
- An example forecast was shown for raw and calibrated disaggregation, and we note that these data are very flexible, where thresholds for extremes can be edited for different use cases, hazards, regional thresholds, etc.
- Verification of the extreme disaggregated forecast wasn't bad! Calibrated disag. was slightly better than raw disag.

Future Work

- While we have focused on 1 variable here, this method can be extended to a multivariate space, and correct or include the covariance between variables (such as temperature and precipitation)
- We have also focused on 1 model, and this method can be extended to the entire suite of North American Multi-Model Ensemble (NMME) models, and to Calibrated, Bridged, and Merged (CBaM) hindcasts/forecasts
- We are currently testing additional thresholds/calculation methods
- Finally, can be extended to real-time forecasts

Thank you!

Acknowledgements: Sarah Strazzo (adaptation of CBaM for CPC); Andrew Schepen and QJ Wang (our partners at CSIRO, CBaM methodology and disaggregation methods); Yawen Shao (BJP Calibration + Trend methodology)
This project funded by the Climate Test Bed (CTB)

References

Schepen A, Everingham Y, Wang QJ (2019) On the Joint Calibration of Multivariate Seasonal Climate Forecasts from GCMs. Mon Wea Rev 148:437–456. https://doi.org/10.1175/MWR-D-19-0046.1

Schepen A, Wang Q, Everingham Y (2016) Calibration, Bridging, and Merging to Improve GCM Seasonal Temperature Forecasts in Australia. Monthly Weather Review 144:. https://doi.org/10.1175/MWR-D-15-0384.1

Schepen A, Wang QJ, Robertson DE (2014) Seasonal Forecasts of Australian Rainfall through Calibration and Bridging of Coupled GCM Outputs. Mon Wea Rev 142:1758–1770. https://doi.org/10.1175/MWR-D-13-00248.1

Schepen A, Everingham Y, Wang QJ (2020) Coupling forecast calibration and data-driven downscaling for generating reliable, high-resolution, multivariate seasonal climate forecast ensembles at multiple sites. International Journal of Climatology 40:2479–2496. https://doi.org/10.1002/joc.6346

Shao Y, Wang QJ, Schepen A, Ryu D (2020) Embedding trend into seasonal temperature forecasts through statistical calibration of GCM outputs. International Journal of Climatology n/a: https://doi.org/10.1002/joc.6788

Strazzo S, Collins DC, Schepen A, et al (2019) Application of a Hybrid Statistical–Dynamical System to Seasonal Prediction of North American Temperature and Precipitation. Mon Wea Rev 147:607–625. https://doi.org/10.1175/MWR-D-18-0156.1

Wang QJ, Schepen A, Robertson DE (2012) Merging Seasonal Rainfall Forecasts from Multiple Statistical Models through Bayesian Model Averaging. J Climate 25:5524–5537. https://doi.org/10.1175/JCLI-D-11-00386.1