Recalibrating HadGEM-GA6 data for South East China

Simon Tett, Nico Freychet, Xian Zhu, Sarah Sparrow & Buwen Dong.

2021-12-03

Summary

We applied the calibration method of Bellprat, Guemas, Doblas-Reyes, and Donat (2019) to data from an event attribution study on temperature & drought for south eastern China in 2020. We calibrate using data from 1961-2013 so the 2020 event allows an out-of-sample test of the method. The calibration does not change our conclusions for temperature indices but does impact our claim of a significant anthropogenic impact on the probability of severe drought. However, this change is largely due to adjustment of the simulated to the observed trend. Further, we find that the calibrated ensemble suggests the 2020 summer temperature were so improbable that we conclude that the method needs to be used very cautiously.

Introduction

In February, the EERCH project ran a successful virtual workshop in event attribution. At this workshop, three events were analysed using data from multiple simulations of the N216 HadGEM-GA6 atmospheric model. They were:

- A cold extreme in late Spring 2020 in Northern China
- Heavy rainfall leading to widespread flooding in the Yangtze valley during June/July 2020
- A combined drought and heatwave in the summer of 2020 in South Eastern China.

One issue of concern for attribution studies is the reliability of the models. Our usual approach has been to evaluate the model in terms of its ability to reproduce the climatological distributions of the variable being used for the attribution study as well as its ability to reproduce the circulation patterns and mechanisms that are believed to drive such events in reality.

Bellprat et al. (2019) argue that the forecast reliability is an important metric and corrections to model distributions from this are appropriate. This approach was criticised by Philip et al. (2020) for two main reasons. First, there is no obvious relationship between forecast skill and changing frequency of events; secondly, that correcting the simulated trend to the observed trend is, in essence, throwing out much useful information from model ensembles. Nevertheless, this report presents a sensitivity study using the approach outlined in Bellprat et al. (2019). Of the three events analysed at the workshop we chose to focus on the combined drought and heatwave event. We do this as three different indicators were considered in the event including a rainfall indicator (see Zheng et al, 2021; attached).

Data & Methods

The data used for the reliability assessment and calibration were the 53-year period from 1961-2013. The HadGEM-GA6 ensemble consists of 15 members for this period and reliable observations for China are available from 1960 onwards (Zhao, Zhu, & Xu, 2014). For risk ratio computations, we make use of two 525-member ensembles for 2020: HistoricalExt and HistoricalNatExt. HistoricalExt is driven with observed sea surface temperatures (SSTs), sea ice concentrations (SICs), and natural & historical forcings while HistoricalNatExt has anthropogenic effects removed from both boundary conditions and forcings. All data was processed during the workshop and we work with average values of June-July-August (JJA) near-surface temperature anomalies (TAS), percentage of anomalous precipitation (PAP) and the anomalous number of hot days (NHD). All averages were taken over land points in South Eastern China (22°–27°N, 111°–120°E) and anomalies were relative to 1981–2010. Observed and simulated data from the Historical, HistoricalExt and HistoricalNatExt ensembles are available.

Reliability

Forecast reliability was assessed using terciles for each variable and forecast probabilities grouped by four equally spaced probability bins. For a reliable forecast we would expect that the fraction of observations that are in the tercile, when that tercile is forecast, should be similar to the forecast probability(Palmer & Weisheimer, 2018). We use a very crude quantisation on both magnitude and probability as we only have 53 data samples. We do not consider uncertainty in our reliability estimates, which is likely considerable.

To help the reader (and writer) better understand this, a simple example is given. For years when the model ensemble has a high probability of being in the upper tercile of the TAS distribution then, if it were reliable, we would expect a large fraction of the observations **in these years** to be in the upper tercile of the TAS distribution.

Calibration

The calibration of simulated data that Bellprat et al. (2019) propose is:

$$y = \alpha x_{EM} + \beta x'_{EM} + \gamma x_{TR}$$

Where y is the calibrated simulated data, x_{EM} the detrended ensemble mean, x'_{EM} the ensemble with the trend and ensemble mean removed, and x_{TR} the ensemble-mean linear trend. The three coefficients are calculated as: $\alpha = |\rho| \frac{\sigma_O}{\sigma_{x_{EM}}}$, $\beta = \sqrt{1 - \rho \frac{\sigma_O}{\sigma_{x'_{\mathrm{EM}}}}}$ and $\gamma = \frac{t_0}{t_x}$.

ρ is the correlation between x_{EM} and the observed detrended timeseries. σ_O is the standard deviation of the detrended observed timeseries. $\sigma_{x_{EM}}$ is the standard deviation of the ensemble-mean timeseries. $\sigma_{x'_{\mathrm{EM}}}$ is the standard-deviation of the detrended intra-ensemble data. t_0 and t_x are the observed and simulated trends.

We use the 15-member Historical ensemble from 1961-2013 to do this calibration and then apply the three coefficients to the HistoricalExt and HistoricalNatExt data in 2020. We extend the simulated 1961-2013 trend to 2020 for the HistoricalExt data. Given the criticism of Philip et al. (2020) we also explore setting $\gamma = 1$ which means no trend correction is made. For HistoricalNatExt we assume a trend of zero. On global scales this is reasonable as natural forcing, with the exception of El Chichon and Pinatubo, varies by only a small amount (IPCC, 2021). If the data from the 15-member Natural ensemble were available we could have computed a trend from this and then applied it to the HistoricalNatExt 2020 data.

The 2020 data is an out of sample calibration. That means we can compare raw and calibrated simulations for 2020 with observations. We can then examine the relative likelihoods for the raw and calibrated cases shedding some light on the relevance of forecast skill calibration to event attribution.

We focus on the probability ratio (PR) of the 2020 event in the Historical world compared to the Natural world. We estimate this by fitting 2020 data to either a normal distribution (TAS & NHD) or to a generalized extreme value distribution (GEV; PAP). From these fits we then compute the probability of TAS or NHD values larger than those observed in 2020 or values of PAP drier than those observed in 2020. We do this for the raw data, the calibrated data and the calibrated data with no trend correction. Thus, for each variable we have three PR values.

Uncertainty estimation

We estimate uncertainty in the Probability Ratio by bootstrapping (Efron & Tibshirani, 1993). We remove the trends from the 1961-2013 observations and ensemble. These residuals are then randomly resampled with replacement and added back onto the trends. From this α, β, γ are computed which in turn are applied to the resampled 2020 HistoricalExt and HistoricalNatExt data. Finally, PR values are computed. This process was repeated 1000 times. From these bootstrap samples a log-normal fit to the data was made and confidence computed. If the log-normal fit was poor, determined using a Kolmogorov-Smirnov (KS) test, then a percentile bootstrap interval (Paciorek, Stone, & Wehner, 2018) was used.

Unless stated otherwise the 5-95% confidence limits are shown. Data and software that computes results and plots figures is available at https://github.com/SimonTett/recalibrate with a GPLv3 license.

Results

First examining the model reliability diagram (Figure 1). For TAS the cold terciles forecasts are on the 1-1 line so we would consider those forecasts reliable. Forecasts for the middle and warm terciles are not so close to the 1-1 line but given the limited amount of data we believe these are reliable. For NHD it appears that forecasts for the highest tercile are unreliable as the number of observations in this tercile are small when the ensemble has a high probability of being in this tercile. However, this may be because there are only a small number of cases when the ensemble has high probability of being in the warm tercile. Finally examining the PAP reliability. Here there are quite large deviations from the 1-1 line particularly for the wet tercile which shows similar behaviour to the high NHD tercile. PAP is very variable with year-to-year variations of 60% (Figure 1 of Zheng et al. (2021)) and there are very few cases where the ensemble has a high probability of any particular tercile. This suggests that sampling error may explain the discrepancies from the 1-1 line.

Turning now to the 2020 temperature indices (Figure 2). For both TAS and NHD, regardless of the use of calibration, the probability of temperatures or number of hot days as observed in the Natural ensemble is below 10^{-5}. We would conclude that such events would never happen in the natural world. Reflecting this, PRs for TAS (Figure 3), regardless of the use of calibration, are very large. Uncertainties are also very large and larger for the calibrated cases likely arising from uncertainty in the calibration parameters. NHD shows similar behaviour to TAS though with smaller PRs. Nevertheless, we conclude that human influences have enormously increased the risk of TAS and NHD values such as occurred in 2020 in South Eastern China.

For PAP which was used as a drought indicator where the raw model ensembles suggest a PR of about two which is significantly different from one, suggesting human influences have increased the risk of drought (Figure 3). The likely driver of this risk increase is aerosols rather than greenhouse gases. Calibrating the ensemble gives a PR of 1.1 with large enough uncertainties that we would conclude that human influences had not impacted the risk of such a damaging drought as observed. Not correcting the trend gives similar results to the raw simulated data. This suggests that the highly uncertain trend correction is driving the PR results.

In the raw ensemble the probability of JJA temperatures as observed (or larger) is about 0.05 (the event is a one-in-twenty occurrence; Figure 2). Calibrating the ensemble distributions results in probabilities less than 10^{-2} regardless whether the trend is corrected or not. The likelihood ratios of the raw ensemble to the fully calibrated ensemble are 26 while, if the trend is not calibrated, the likelihood ratio is 9 (Table 1). Based on this, out of sample, test then the calibration approach of Bellprat et al. (2019) makes the model distributions much worse for TAS. For NHD and PAP calibration does not dramatically change the probability of the observed event. Given our finding for TAS and the issues raised by Philip et al. (2020) we do not think the recalibrated results, particularly when the ensemble trends are adjusted to the observed trends, are reliable and so should not be used. This is particularly true for rainfall and temperatures in Eastern China which shows considerable decadal variability due to Maiyu variability (Ding, Liang, Liu, & Zhang, 2020).

Conclusions

Calibration of the model ensemble leads to little difference in the original findings for TAS and NHD. PR's are large enough and calibration does not change the Natural distribution enough that the conclusions change. Such a hot summer would not occur in the Natural world.

Calibration of the model ensemble does impact PAP. In particular correcting the trend removes the claim that risk of drought event was roughly doubled by anthropogenic drivers. Without trend correction then risk of drought roughly doubles. This suggests it is important to understand the trend in simulated rainfall in current climate models over eastern China as adjustment of this trend changes the conclusion.

Reliability calibration produces, at least for TAS, distributions which suggest that the 2020 event was about a 1-in-500 year event compared to the raw model ensemble which suggests it is a 1-in-20 year event. Even with sampling bias one concludes that the raw model ensemble is much more plausible than the calibrated model. Given this and the criticisms of Philip et al. (2020) we think the use of methods like Bellprat et al. (2019) which use forecast reliability to correct model distributions should be used very cautiously.

Figures & Tables

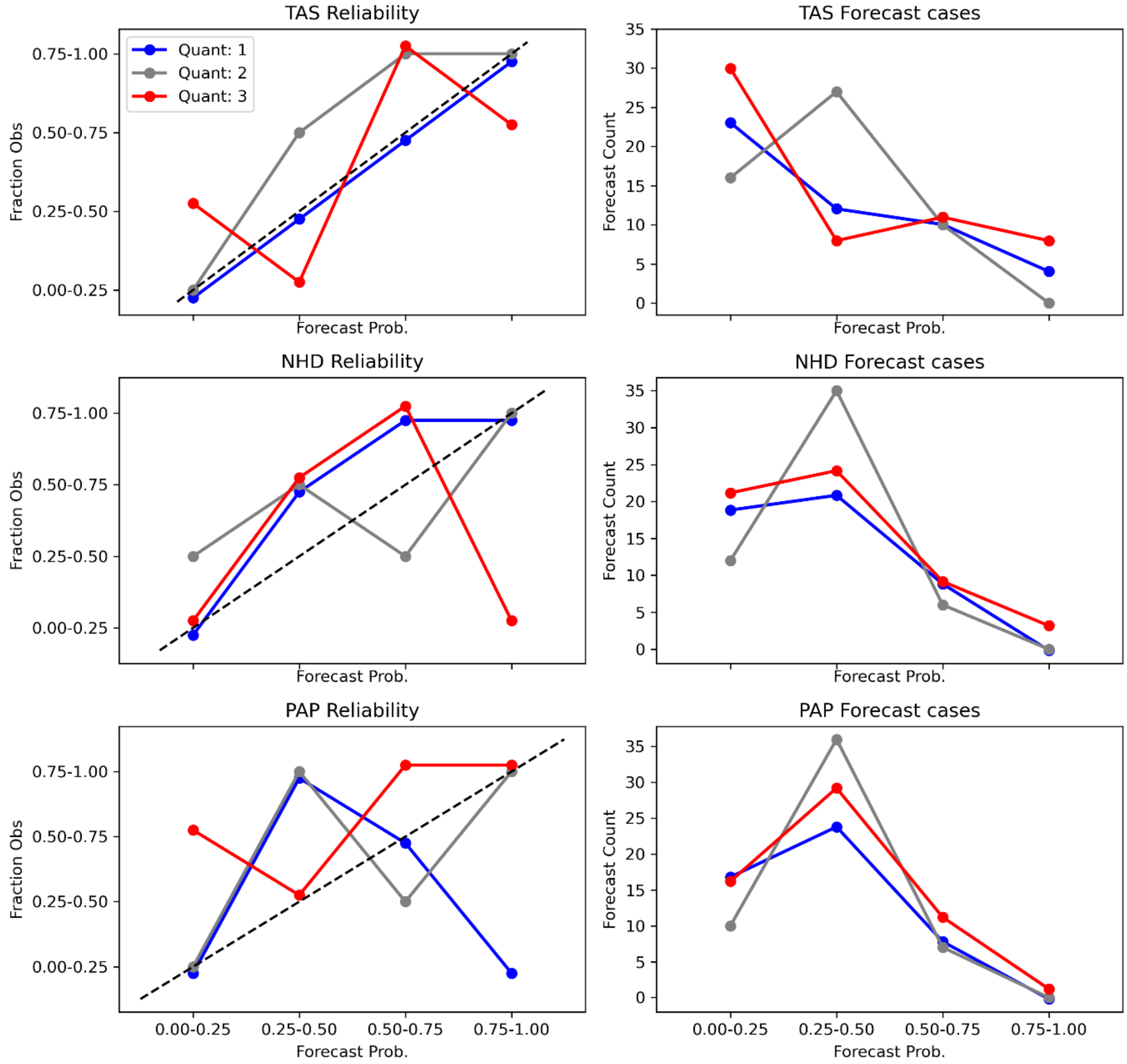

Figure 1: Reliability diagram. The left-hand column shows for the three terciles (#1 in blue, #2 in grey & #3 in red) for each of the three variables. The x-axis shows the probability interval for each forecast while the y-axis shows the fractional interval of observations that occurred for those forecast cases. For a perfectly reliable model we would expect the values to be on the 1-1 line (dashed black line). The right-hand columns show the number of years in each forecast probability range, for each tercile.

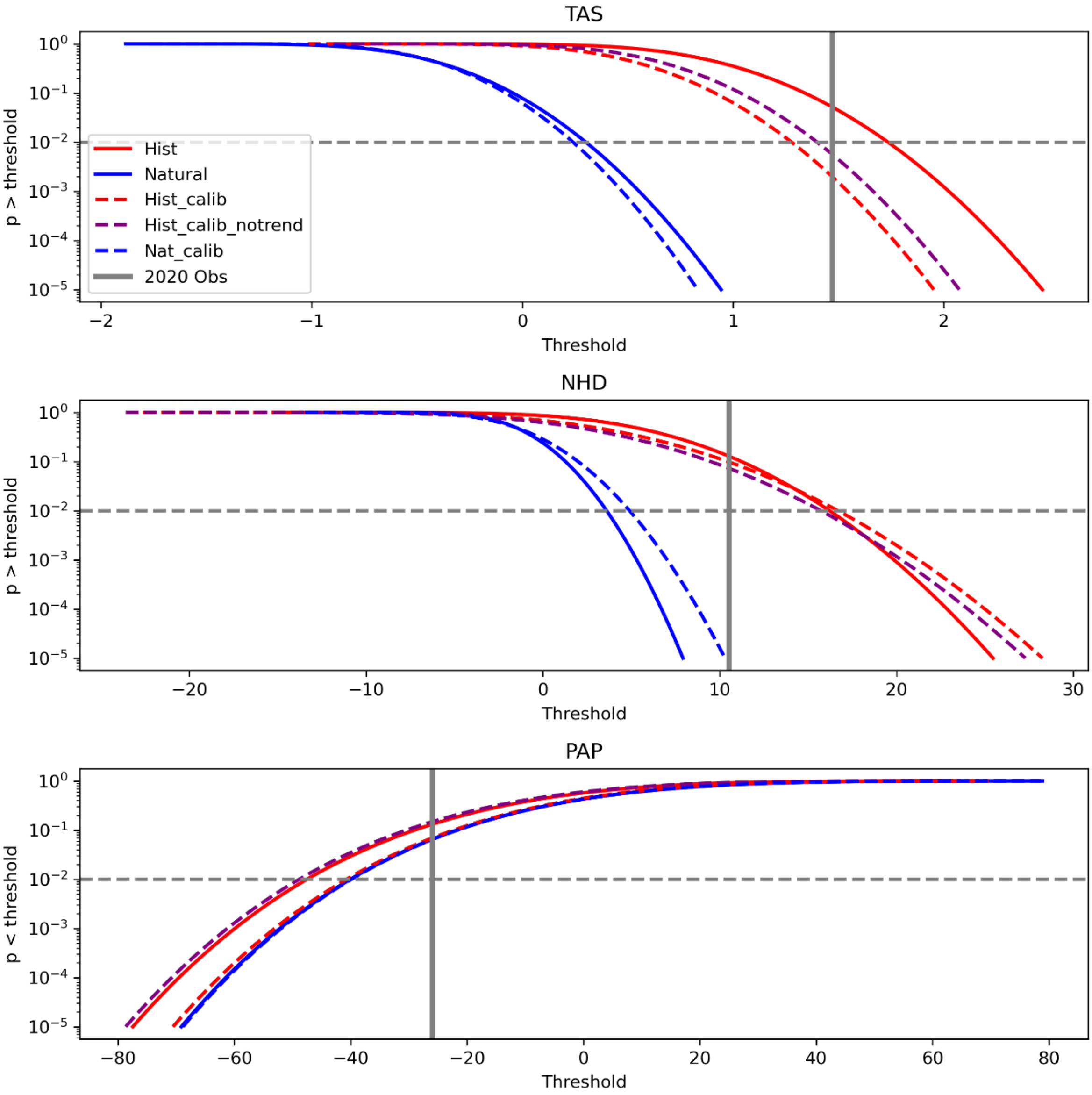

Figure 2: Cumulative probability (y-axis) of exceeding threshold (x-axis) for TAS (top) & NHD (middle). Bottom shows PAP where probability shown is probability of being below a threshold. Y-axis is a logarithmic axis. Key in top plot identifies coloured/dashed lines. +The horizontal dashed grey line denotes a probability of 0.01. Hist 2020 is the distribution from the un-calibrated HistoricalExt summer 2020 HadGEM-GA6 ensemble, Nat 2020 is the distribution from the un-calibrated HistoricalNatExt summer 2020 HadGEM-GA6 ensemble. "Calib" shows PDFs where the data is calibrated and "Calib (no trend)" where the calibration does not affect the trend. The vertical grey line shows the observed values for 2020.

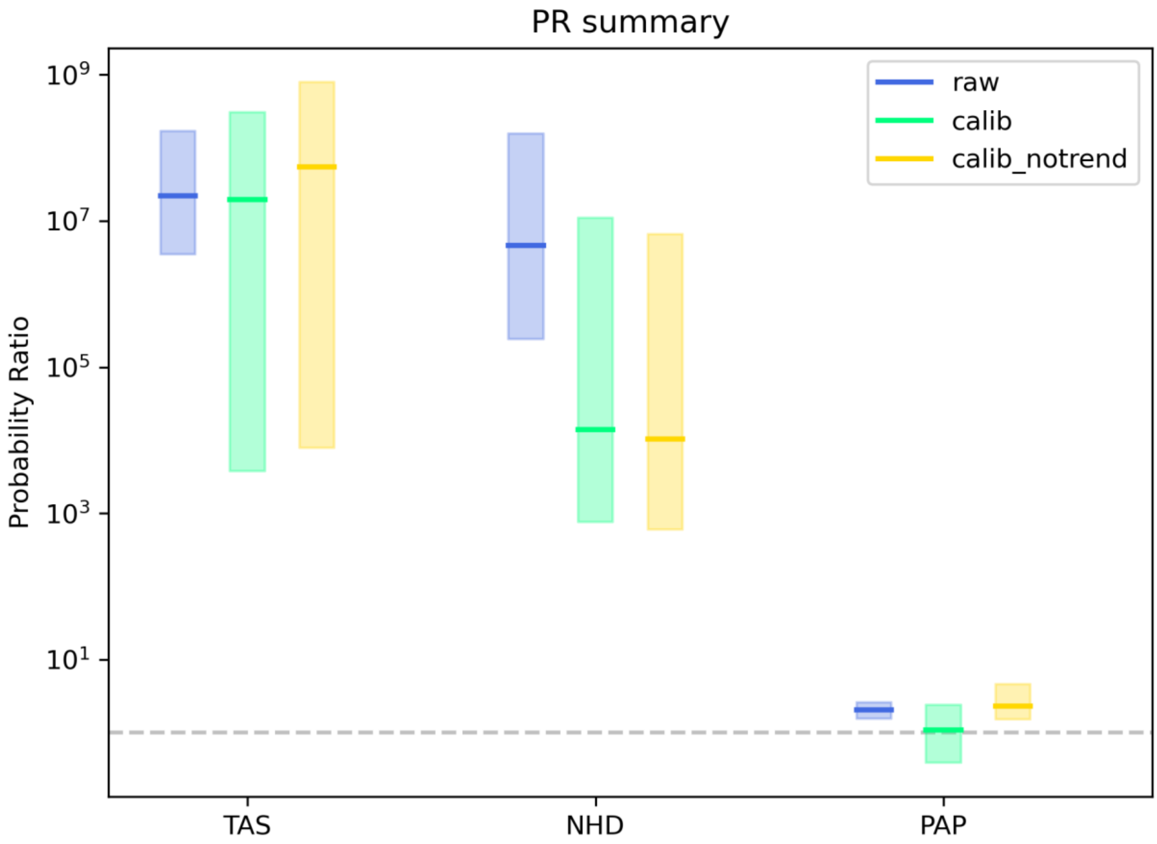

Figure 3: Probability ratios (PRs; y-axis) for three variables (x-axis) with raw (blue), calibrated (green) and calibrated with no trend correction (gold). PR ratio uses a log scale. For each variable & calibration analysis the best estimate (horizontal line), and 5-95% confidence range (bar) is shown. Dashed horizontal line shows PR=1. Uncertainty estimated from a log-normal fit to the bootstraped data. If this log-normal fit is ***not*** *adequate, using a Kolmogorov-Smirnoff test a the 10% level, then the percentile bootstrap uncertainty range (on logs of PR) is shown. The log-normal fit for the Calibrated TAS and NHD data fails.*

Table 1: Probabilities of event in 2020 HistoricalExt ensemble and likelihood ratios for raw to calibrated ensembles.

Variable	Probability of event in raw ensemble	Likelihood ratio of Raw to calibrated	Likelihood ratio of Raw to calibrated - notrend
TAS	0.052	26.1	9.2
NHD	0.13	1.3	1.8
PAP	0.13	1.9	0.9

References

Bellprat, O., Guemas, V., Doblas-Reyes, F., & Donat, M. G. (2019). Towards reliable extreme weather and climate event attribution. *Nature Communications, 10*(1), 1732. doi:10.1038/s41467-019-09729-2

Ding, Y., Liang, P., Liu, Y., & Zhang, Y. (2020). Multiscale Variability of Meiyu and Its Prediction: A New Review. *Journal of Geophysical Research: Atmospheres, 125*(7). doi:10.1029/2019jd031496

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap* (Vol. 57). New York: Chapman and Hall.

IPCC. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*: Cambridge University Press.

Paciorek, C. J., Stone, D. A., & Wehner, M. F. (2018). Quantifying statistical uncertainty in the attribution of human influence on severe weather. *Weather and Climate Extremes, 20*, 69-80. doi:10.1016/j.wace.2018.01.002

Palmer, T. N., & Weisheimer, A. (2018). A Simple Pedagogical Model Linking Initial-Value Reliability with Trustworthiness in the Forced Climate Response. *Bulletin of the American Meteorological Society, 99*(3), 605-614. doi:10.1175/bams-d-16-0240.1

Philip, S., Kew, S., van Oldenborgh, G. J., Otto, F., Vautard, R., van der Wiel, K., . . . van Aalst, M. (2020). A protocol for probabilistic extreme event attribution analyses. *Advances in Statistical Climatology, Meteorology and Oceanography, 6*(2), 177-203. doi:10.5194/ascmo-6-177-2020

Zhao, Y., Zhu, J., & Xu, Y. (2014). Establishment and assessment of the grid precipitation datasets in China for recent 50 years. *Journal of the Meteorological Sciences, 34*(4), 414-420. doi:10.3969/2013jms.0008

Zheng, Z., Wang, K., Bu, L., Wang, Y., Li, H., Zhu, X., . . . Nanding, N. (2021). Anthropogenic influences on the extremely dry and hot summer of 2020 in Southern China. *Bull. Am. Meteorol. Soc.* (Submitted)