CSIRO Large Ensemble Assessment Project

Model bias in assessment of extreme events

Damien Irving, Annette Stellema, James Risbey, Didier Monselesan, Tess Parker, Carly Tozer

WCRP EPESC/LEADER workshop, Busan

18th July 2025

Questions

For a given extreme event

- How much should we trust model assessments of event likelihood?

Questions

For a given extreme event

- How much should we trust model assessments of event likelihood?

Rephrase

- what is the role of model errors in simulating extremes?

- should we bias-correct the models?
- if so, how should we perform the bias corrections?
  - 'soft' vs 'hard' bias correction
- what is the role of bias correction on likelihood assessments?
- what is the role of model selection on likelihood assessments?

UNSEEN

model likelihoods for extreme events

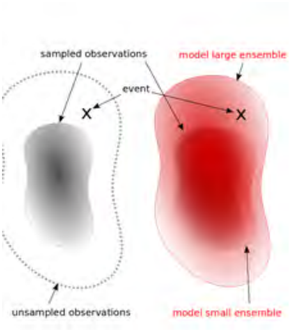
statistics of rare events — pioneered by UKMO analysis
2014 floods England

An extreme event occurs that is hitherto UNSEEN in observations

- large ensembles simulate weather many times over
- provides a much larger sample than observations
- large sample includes many extreme outcomes
- provides a way to assess the likelihood of extremes

An unprecedented extreme event '**X**' occurs

- the event '**X**' is well out of the historical distribution
- observational data presents a vast undersample
- can use Extreme Value Theory (EVT) to learn more about the tail
- but if the event is way off-scale, then EVT mostly reflects the assumptions you put in
- models can provide large samples
- but are they any good?
  - statistical tests
  - process studies



Large ensemble models

Decadal Climate Prediction Project models

Model	Samples	Members	Runtime	Initialisation		Base Period	Grid
				Month	Years		
CAFE	34,944	96	10	May & Nov	1995-2020	2004-2020	2.0°× 2.5°
BCC-CSM2-MR	3,888	8	10	Jan	1961-2014	1970-2014	1.125°× 1.125°
CanESM5	7,980	20	10	Jan	1961-2017	1970-2017	2.8°× 2.8°
CMCC-CM2-SR5	7,200	10	10	Nov	1960-2019	1970-2019	0.94°× 1.25°
EC-Earth3	6,960	15	10	Nov	1960-2017	1970-2017	0.8°× 0.7°
IPSL-CM6A-LR	4,560	10	10	Jan	1961-2017	1970-2017	1.27°× 2.5°
MIROC6	4,130	10	10	Nov	1960-2018	1970-2018	1.4°× 1.4°
MPI-ESM1-2-HR	5,310	10	10	Nov	1960-2018	1970-2018	0.94°× 0.94°
MRI-ESM2-0	2,400	10	5	Nov	1960-2019	1965-2019	1.125°× 1.125°
NorCPM1	9,440	20	10	Oct	1960-2018	1970-2018	1.9°× 2.5°

Events

extreme events — based here on block maxima

- TXx – hottest day of the year
- RX1 – wettest day of the year

initial tests

- test for and remove any non-independent lead times
- test for distribution dependence on lead time – so we can pool leads
- test for stationarity with calender time – so we can pool decades
- test for fidelity of model distribution with observed distribution (moments tests and KS test)
  - raw (uncorrected) model output
  - bias corrected model output

Bias correction

'hard' bias correction

- e.g. quantile – quantile : $x_{qq} = F_{obs}^{-1}(F_{mod}(x_{mod}))$
- maps the model into the observed distribution
- (somewhat) defeats the purpose of using the model

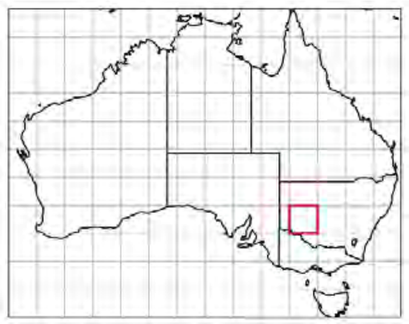
'soft' bias correction

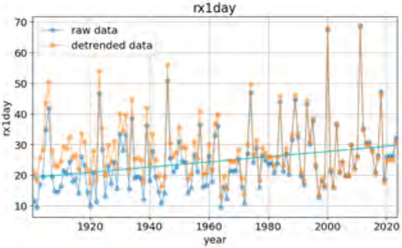
- simple additive : $x_{add} = x_{mod} - (\langle x_{mod} \rangle - \langle x_{obs} \rangle)$ TXx
- simple multiplicative : $x_{mul} = x_{mod}(\langle x_{obs} \rangle / \langle x_{mod} \rangle)$ RX1

'hard' and 'soft' bias correction used here as a crude bounds on the role of bias correction

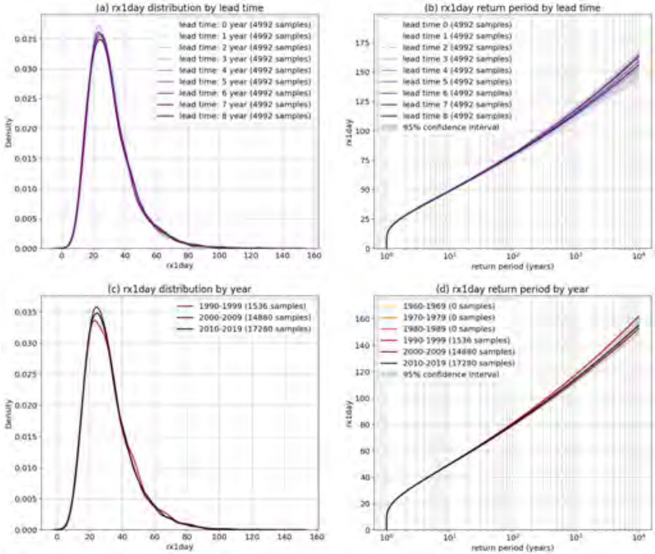
Apply tests to gridded model outputs

RX1day – AGCD gridded observations

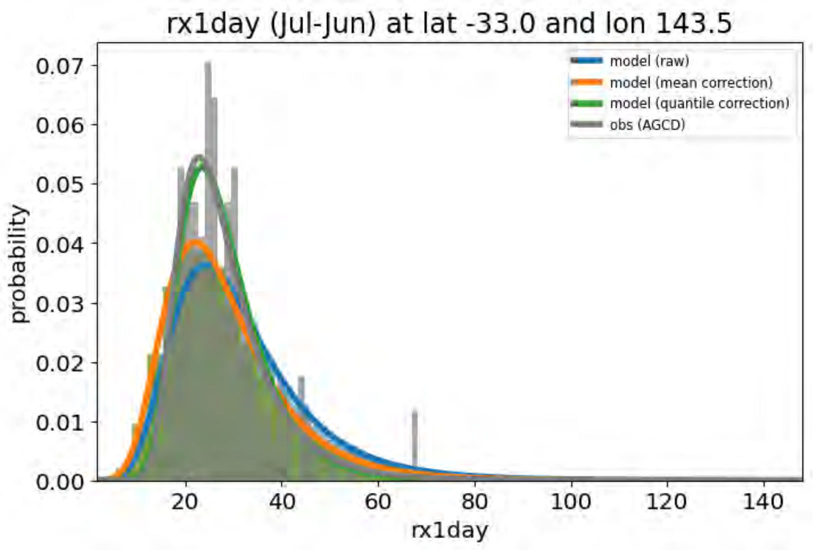




CAFE model : lead time and decade

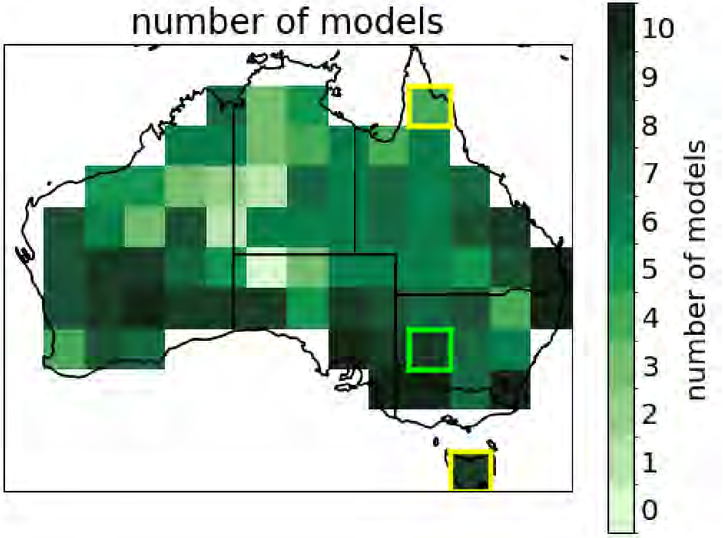


CAFE model : distributions



Models passing KS test

RX1d After multiplicative correction



Account for uncertainties

Sources of uncertainty addressed

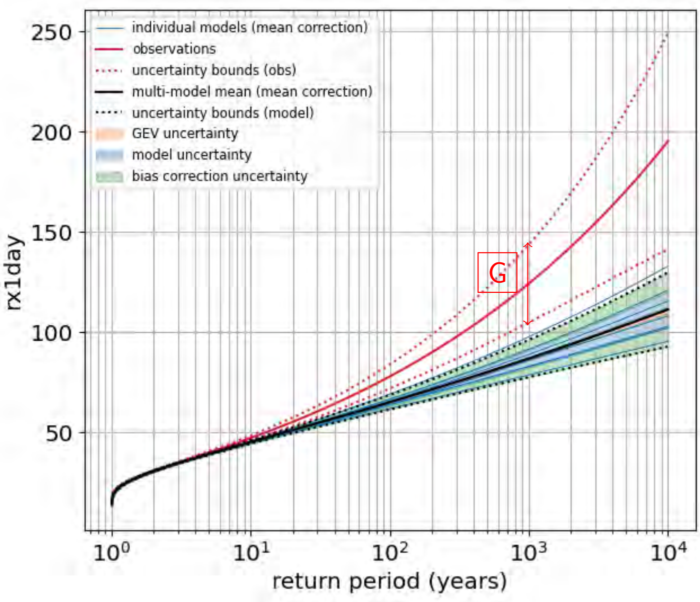
GEV fit, G	Monte Carlo with 1000 resamples of soft bias correction case, variance G^2 over resamples, mean over models for multimodel
bias correction, B	variance B^2 over bias correction methods, $x_{mul} - x_{qq}$, mean over models for multimodel
model selection, M	DCPP multi model spread based on soft bias correction; variance M^2 over models

Adding in the uncertainties

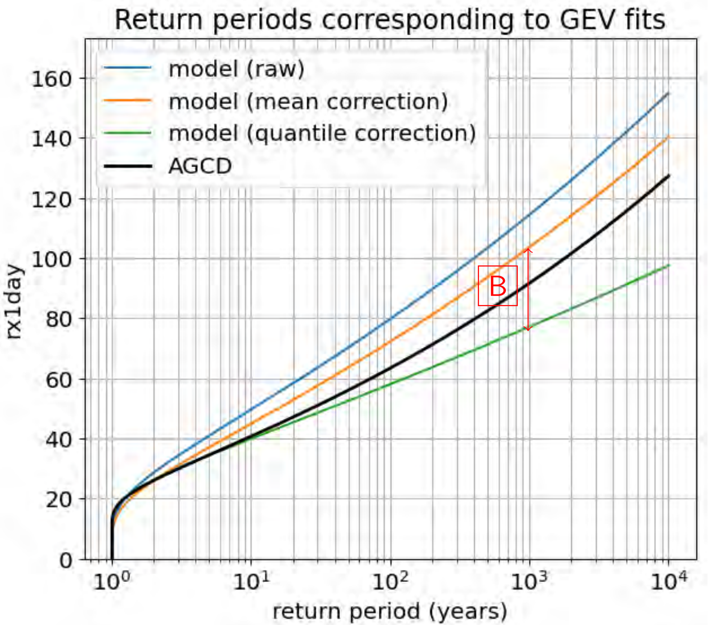
GEV uncertainty, G, model uncertainty, M, bias uncertainty, B

We show the proportion of the total standard deviation due to each type of uncertainty. This has been estimated by considering that the total variance in the return curve (T^2) is the sum of the variance due to GEV uncertainty (G^2), model uncertainty (M^2) and bias correction uncertainty (B^2).
$T^2 = G^2 + M^2 + B^2$
When considering the total standard deviation, T, we would like,
$T = G' + M' + B' = \frac{G}{F} + \frac{M}{F} + \frac{B}{F}$
where the primes denote scaled versions of G, M and B. The common scaling factor, F, is then,
$F = \frac{G+M+B}{T}$
and the boundaries between the different coloured sections in Figure X are at $\pm\frac{G}{F}$, $\pm\frac{G+M}{F}$, and $\pm\frac{G+M+B}{F}$.
$G^2 =$ variance of all the parametric bootstrapped GEV fits (take average of G^2 from each model for ensemble value)
$M^2 =$ variance in the bias corrected (mean scaling) curve for each model (N/A for single model plot)
$B^2 =$ variance in the bc-mean and bc-quantile curves (take average of B^2 from each model for ensemble value)

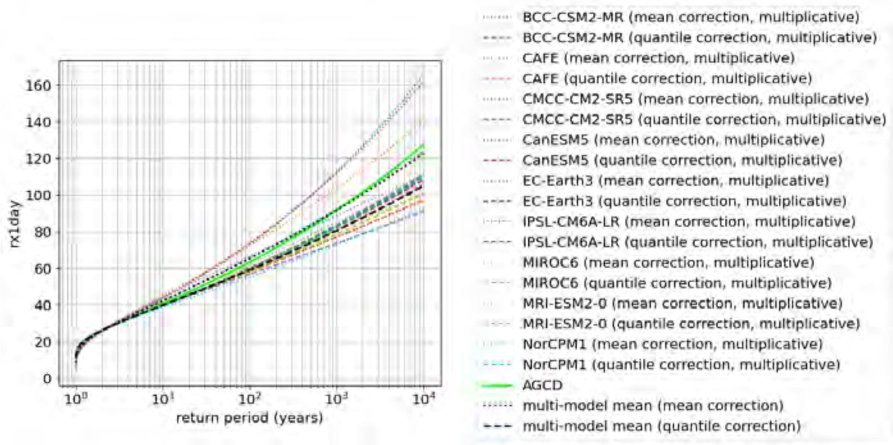
GEV uncertainty : G



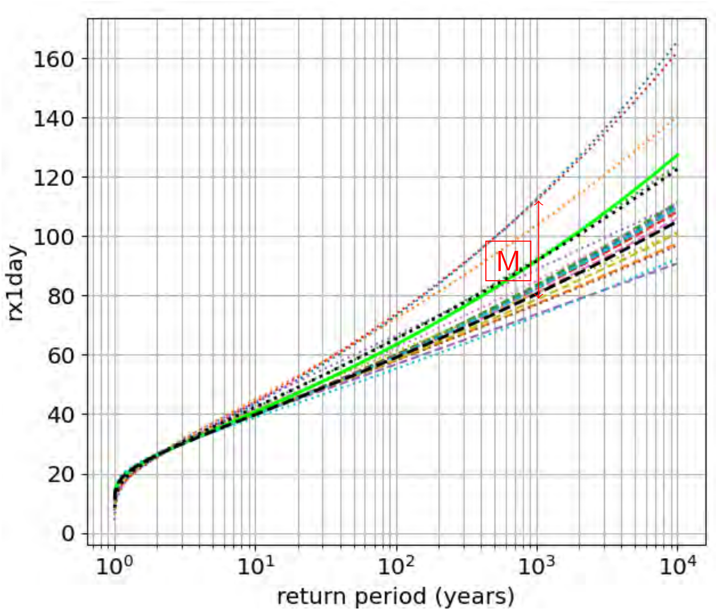
CAFE model : bias correction uncertainty : B



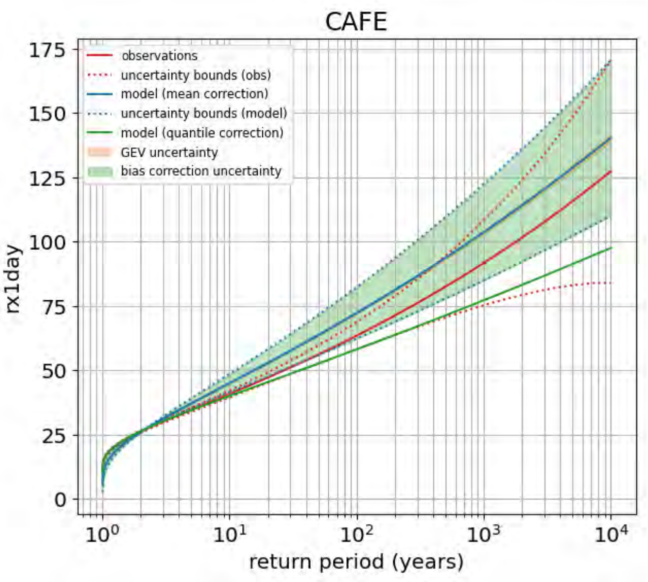
Multi model : model selection uncertainty : M



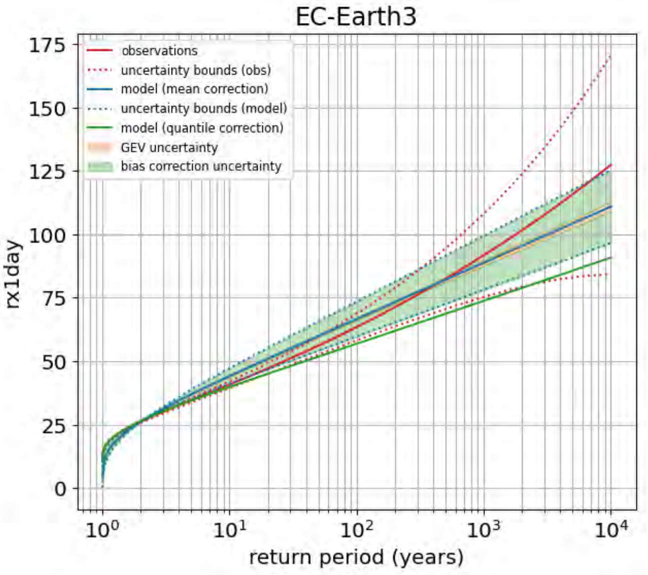
Multi model : model selection uncertainty : M



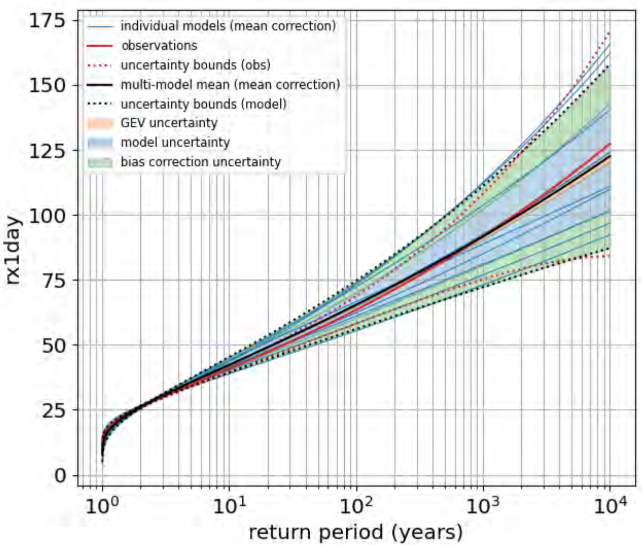
CAFE model : return periods with spread : G + B



EC-Earth model : return periods with spread : G + B

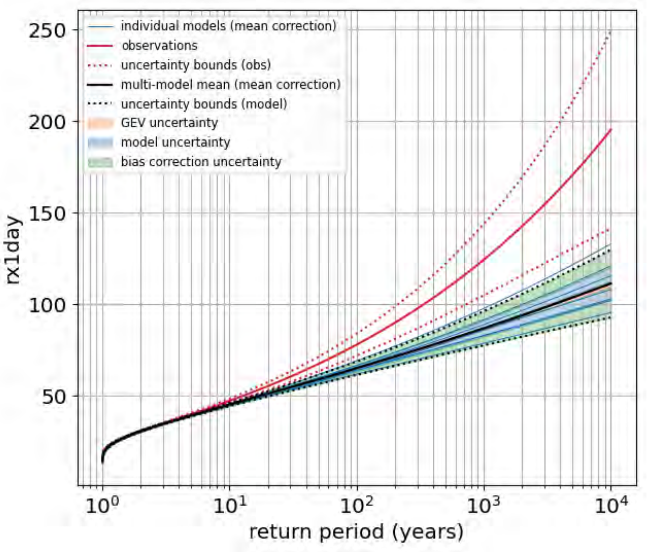


Multi model : return periods with spread : G + B + M

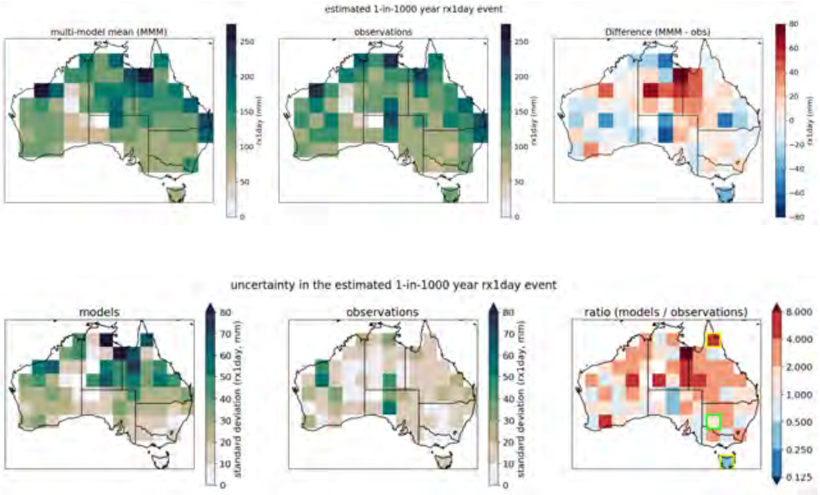


Tasmania grid box

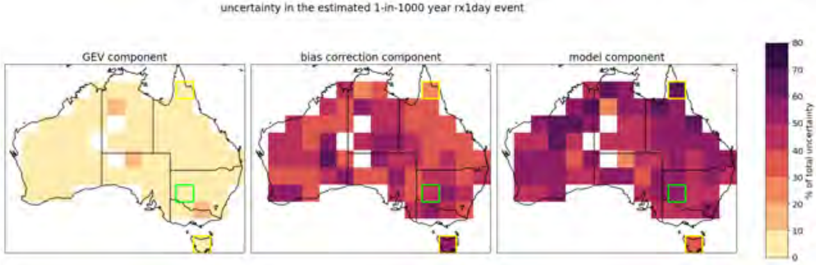
Multi model : return periods with spread : Tas



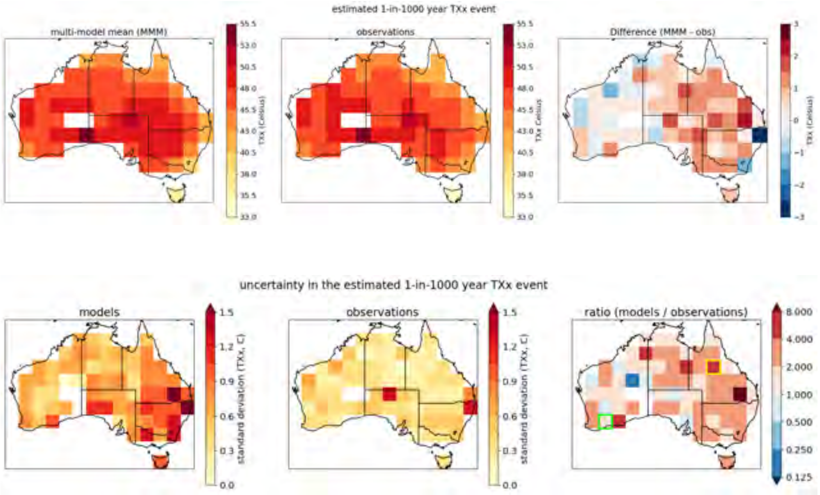
1 in 1000 year RX1d



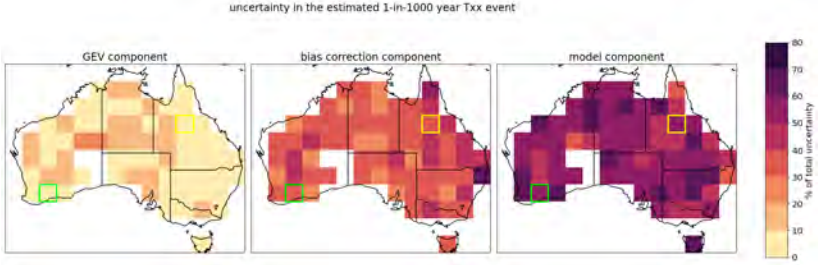
RX1d Components of uncertainty G B M



1 in 1000 year TXx



TXx Components of uncertainty G B M



Model assessment of extreme event likelihoods

RX1, TXx in DCPP

- provide (crude) estimates of uncertainty due to:
  - choice of model
    - model selection differences most important
    - single model estimates underplay uncertainties
    - model spread $\neq$ model error
  - bias correction method
    - also important component of uncertainty
    - need systematic tests of multiple methods
  - GEV uncertainty
    - GEV sampling $\neq$ distribution uncertainty
    - no attempt to account for obs errors
- repeat in higher resolution models

Large Ensemble Assessment Project papers

- Squire, D., D. Richardson, J. Risbey, A. Black, V. Kitsios, R. Matear, D. Monselesan, T. Moore, and C. Tozer 2021: Likelihood of unprecedented drought and fire weather during Australia's 2019 megafires. *npj Clim. Atmos. Sci.*, **4**, 64
- Risbey, J., D. Irving, D. Squire, R. Matear, D. Monselesan, M. Pook, N. Ramesh, D. Richardson, and C. Tozer 2023: A large ensemble illustration of how record-shattering heat records can endure. *Env. Res. Climate*, **2** (3), 1–18
- Irving, D., J. Risbey, D. Squire, R. Matear, C. Tozer, D. Monselesan, N. Ramesh, P. Reddy, and M. Freund 2024: A multi-model likelihood analysis of unprecedented extreme rainfall along the east coast of Australia. *Meteorological Applications* **31** (3), 1–14
- Stellema A. and others 2025: A soft record analysis of extreme heat across Australia. *Meteor. Appl.* In review.
- **Irving, D., A. Stellema, J. Risbey, D. Monselesan, T. Parker, and C. Tozer 2025: UNSEEN uncertainty. In preparation**
- Tozer C. and others: Large ensemble assessment of drought periods in the marginal grain production regions of Australia. In preparation.