Seasonal Forecasting: probabilities, skill, reliability

Some words with some useful pictures

Philip Bett

SECLI-FIRM 4th Stakeholder Workshop 15th June 2020

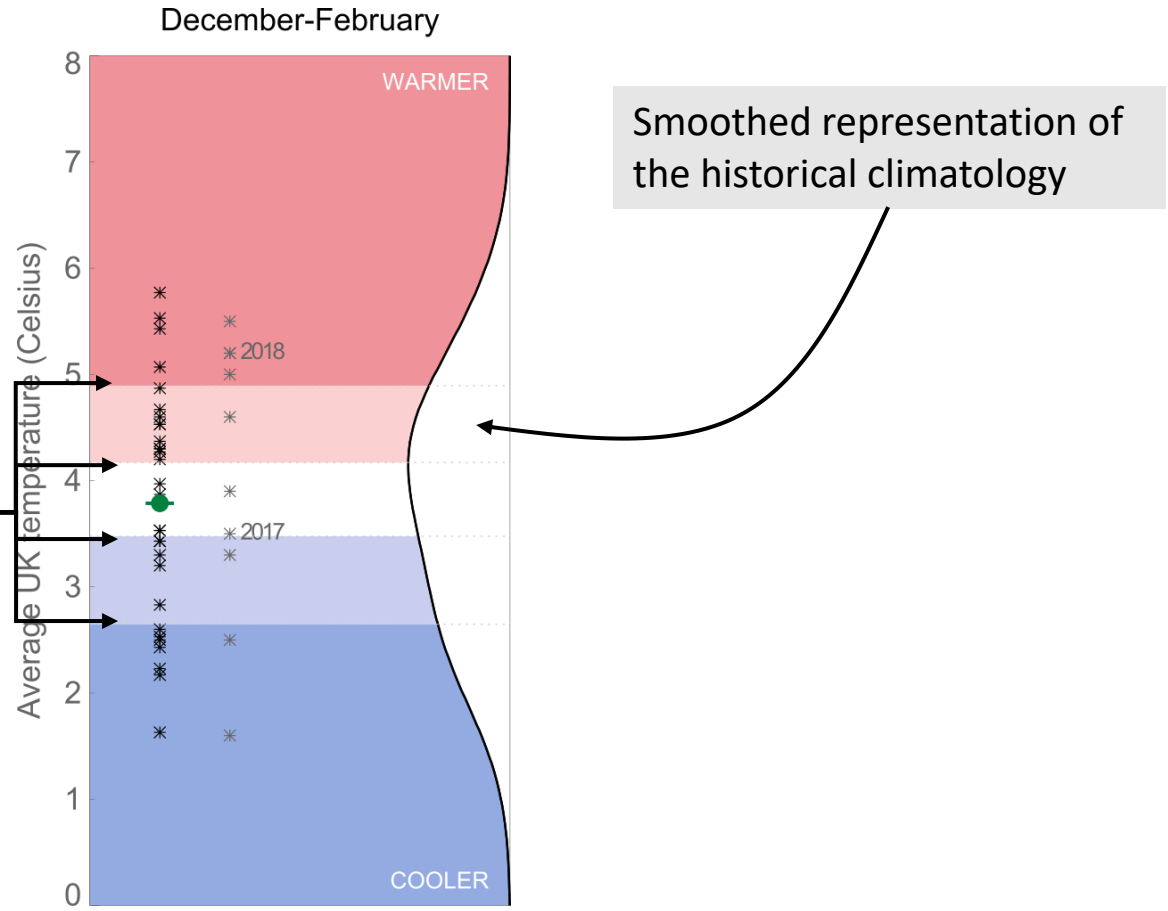
Grant Agreement n. 776868

Probabilistic seasonal forecasts

Seasonal forecasts have to be used probabilistically. They are produced as **ensembles** of forecasts.

These can be compared to the range of historical observations (the observed climatology)

Quintiles: Dividing the 1981-2010 climatology into 5ths (6 years)



Smoothed representation of the historical climatology

- 1981-2010 average
- ∗ Observations 1981-2010
- ∗ Observations 2009-2018

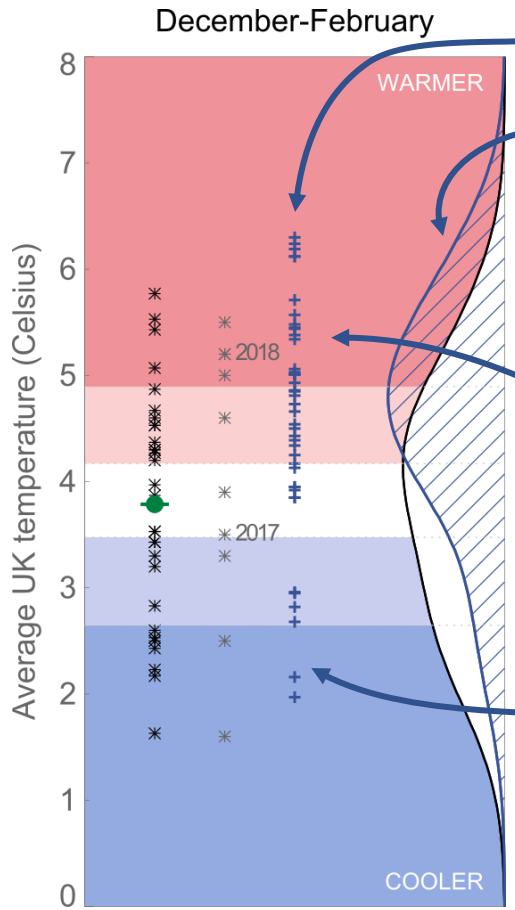
https://www.metoffice.gov.uk/services/government/contingency-planners/index

Grant Agreement n. 776868

Probabilistic seasonal forecasts

Seasonal forecasts have to be used probabilistically. They are produced as **ensembles** of forecasts.

These can be compared to the range of historical observations (the observed climatology)



Forecast ensemble

Smoothed representation of the ensemble as a probability density function (PDF)

Enhanced probability of warmer-than-normal winter

Colder-than-normal winter remains possible, but less likely than an average year.

- 1981-2010 average
- ∗ Observations 1981-2010
- ∗ Observations 2009-2018
- \+ Forecast for 2019 DJF

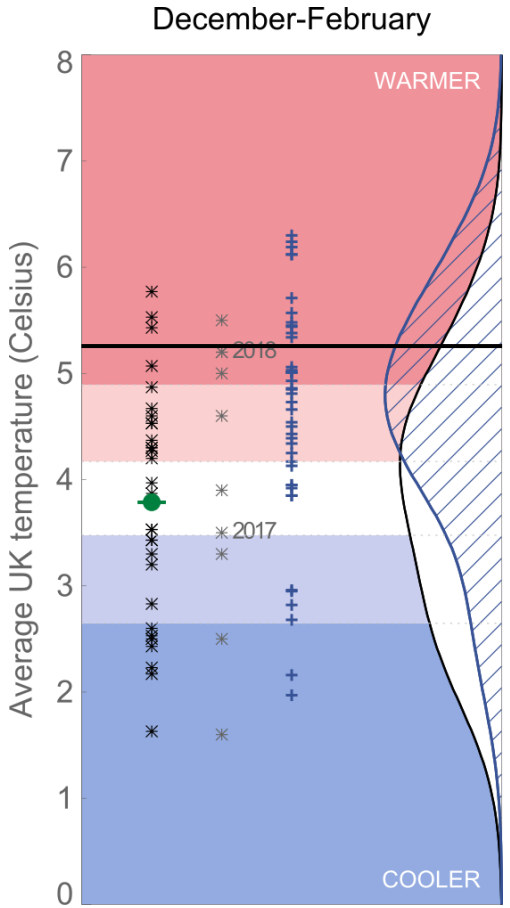
https://www.metoffice.gov.uk/services/government/contingency-planners/index

Grant Agreement n. 776868

Probabilistic seasonal forecasts

Seasonal forecasts have to be used probabilistically. They are produced as **ensembles** of forecasts.

These can be compared to the range of historical observations (the observed climatology)



Actual DJF-mean temperature 👍

- 1981-2010 average
- Observations 1981-2010
- Observations 2009-2018
- Forecast for 2019 DJF

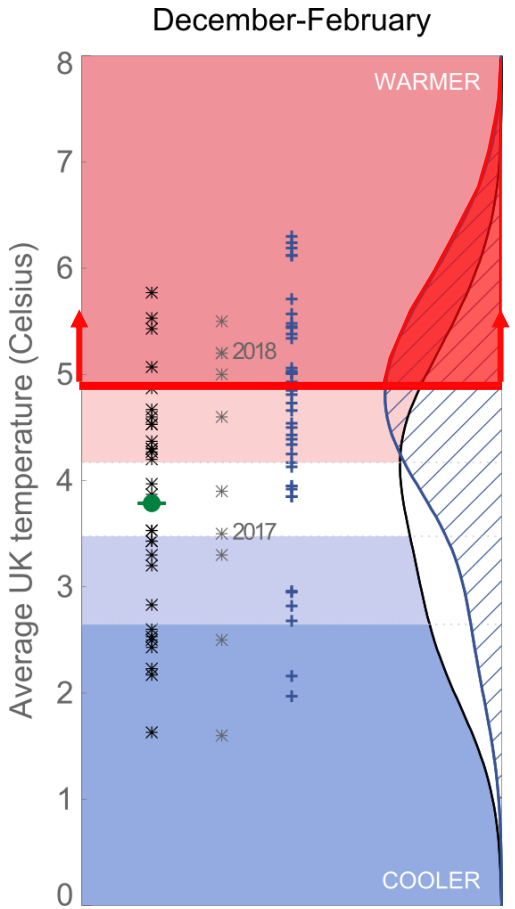
https://www.metoffice.gov.uk/services/government/contingency-planners/index

Grant Agreement n. 776868

Probabilistic seasonal forecasts

Seasonal forecasts have to be used probabilistically. They are produced as **ensembles** of forecasts.

These can be compared to the range of historical observations (the observed climatology)



● 1981-2010 average
∗ Observations 1981-2010
∗ Observations 2009-2018
\+ Forecast for 2019 DJF

Can use **thresholds** to define **events** of interest for decision-making (yes/no):
→ Calculate probability of event occurring

BUT: 🤔

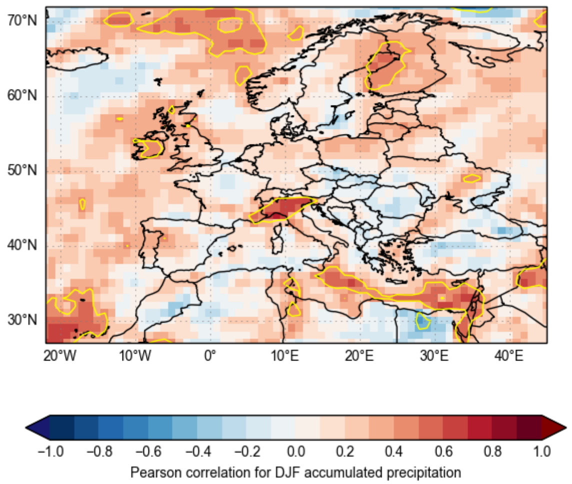
- How to use in a downstream model that requires temperatures?
  - Run model multiple times? Or…?
- What about when our usage is continuous (more/less), not a yes/no decision?

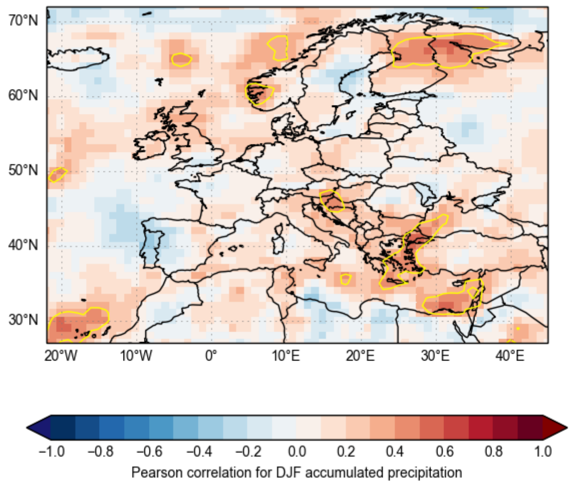
https://www.metoffice.gov.uk/services/government/contingency-planners/index

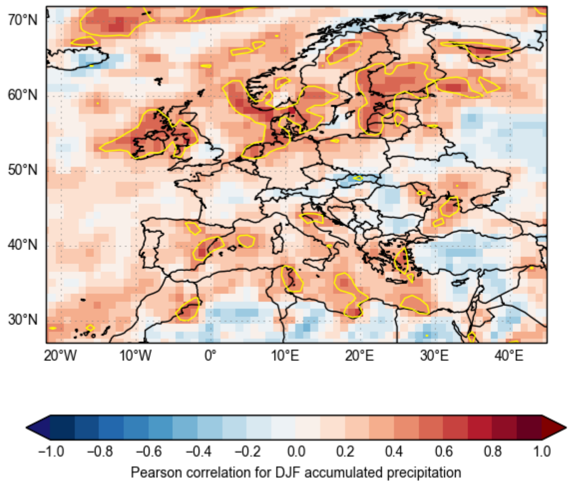
Grant Agreement n. 776868

Skill of seasonal forecasting systems

- Property of a **forecast system**, not a single forecast. Requires “**hindcasts**” over historical period – AND **observations**
- Lots of ways of measuring – not surveying skill scores here!
- For seasonal forecasts in Europe, skill is often not high, and is quite variable/noisy (but still useful!)
- Skill is **estimated** from hindcasts & observations. These cover a limited number of years → **skill is uncertain!**





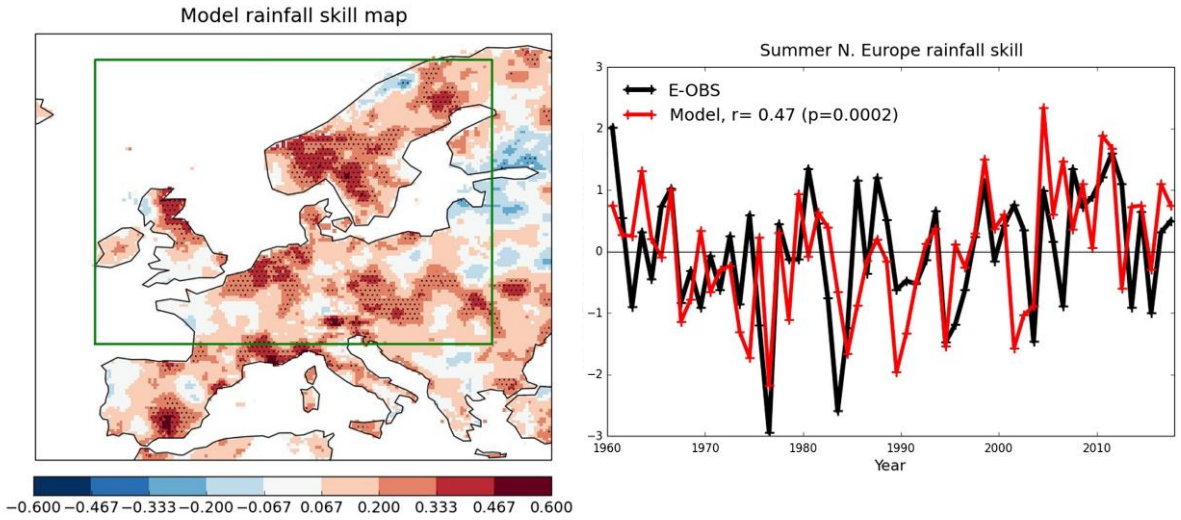


“Significant skill” = “significantly different to zero” i.e. we can be reasonably sure there is non-zero skill!

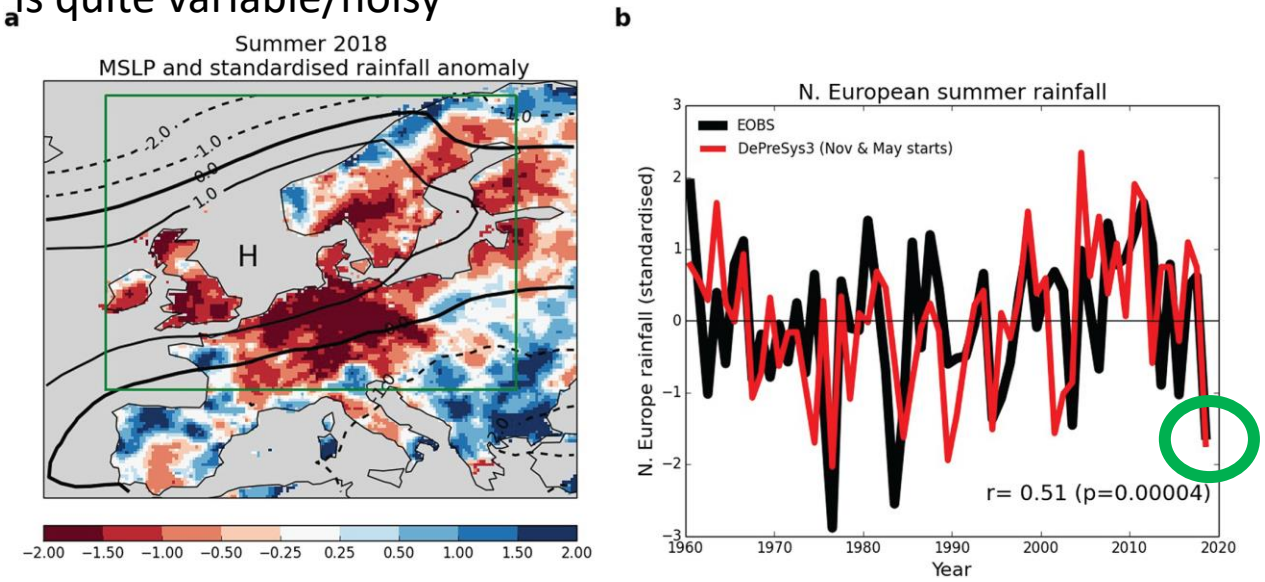
Grant Agreement n. 776868

Skill of seasonal forecasting systems

- Property of a **forecast system**, not a single forecast
- Lots of ways of measuring – not surveying skill scores here!
- For seasonal forecasts in Europe, skill is often not high, and is quite variable/noisy



Dunstone et al. (2018) https://doi.org/10.1002/2017GL076337



Dunstone et al. (2019) https://doi.org/10.1029/2019GL084659

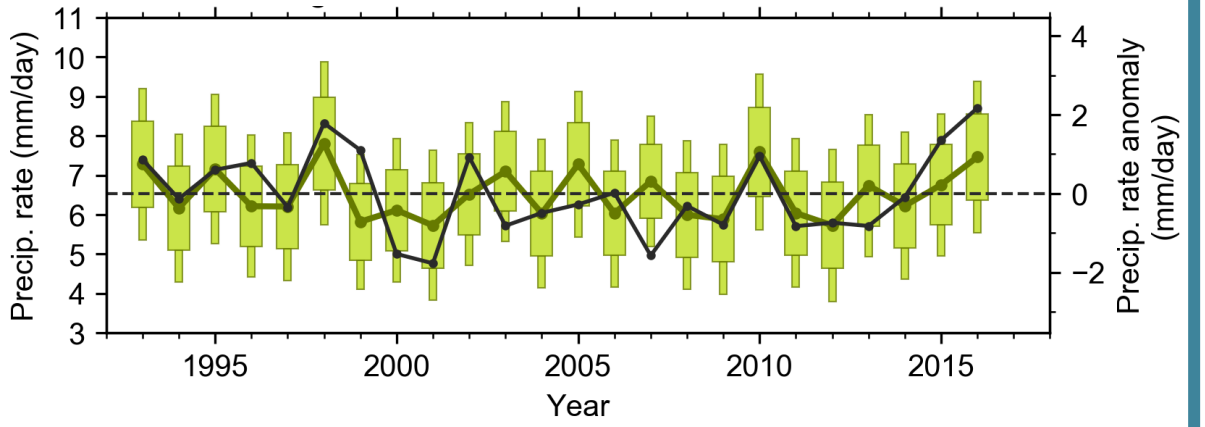
"Significant skill" = "significantly different to zero" i.e. we can be reasonably sure there is non-zero skill!

Grant Agreement n. 776868

Skill of seasonal forecasting systems

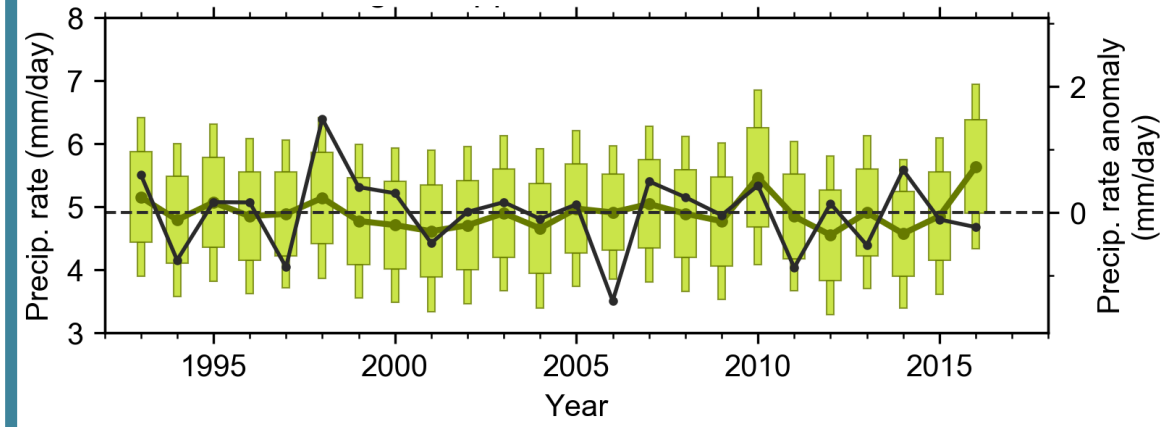
- Property of a **forecast system**, not a single forecast
- Lots of ways of measuring – not surveying skill scores here!
- For seasonal forecasts in Europe, skill is often not high, and is quite variable/noisy

Good skill: r = 0.54, CRPSS = 0.42



Note: Ensemble mean is not always the right sign

Poor skill: r = 0.16, CRPSS = 0.23



Forecasts not responding to drivers

"Significant skill" = "significantly different to zero" i.e. we can be reasonably sure there is non-zero skill!

Grant Agreement n. 776868

Reliability and probabilistic calibration

"Reliability" = how well the forecast probabilities match observed frequencies.

Can assess using a **reliability diagram**,
defined for the **particular event** we're trying to forecast.

This example:

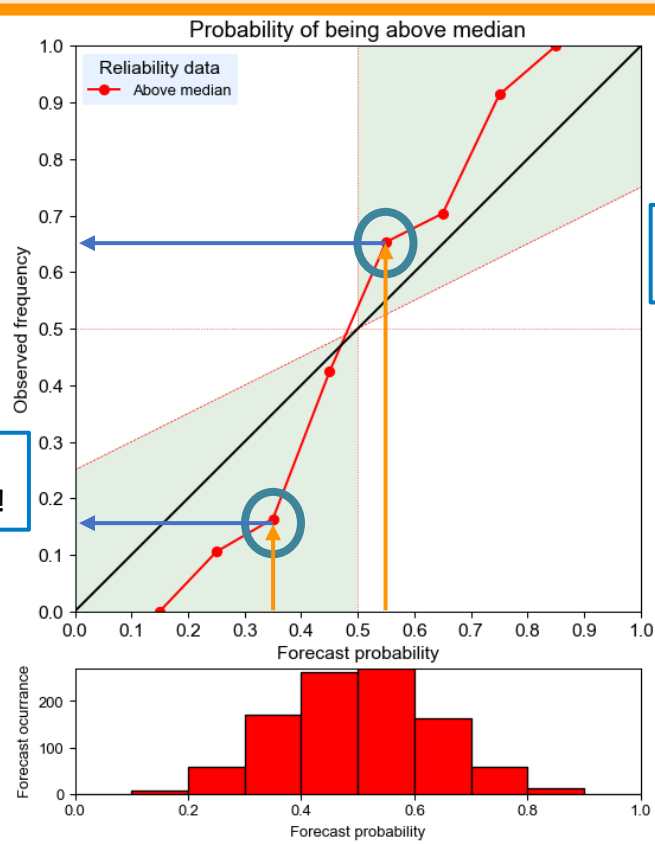
- Is the quantity (e.g. temperature T)
  above the median T_{med} of the historical observations?

Each forecast = a probability that $T > T_{med}$.
Over all observations, this is 50/50!

For a set of probability steps,
Pick all the times a forecast gave that probability,
And for those cases, count how often it actually occurred…

This is **underconfident**:

- Low probabilities aren't low enough
- High probabilities aren't high enough



It actually only occurs 10% - 20% of the time!

When the model forecasts the event with probabilities of 30% - 40%

It actually occurs 60% - 70% of the time!

When the model forecasts the event with probabilities of 50% - 60%

Most forecasts have probabilities around the climatological frequency (50%)

Grant Agreement n. 776868

Reliability and probabilistic calibration

"Reliability" = how well the forecast probabilities match observed frequencies.

Can assess using a **reliability diagram**,
defined for the **particular event** we're trying to forecast.

This example:

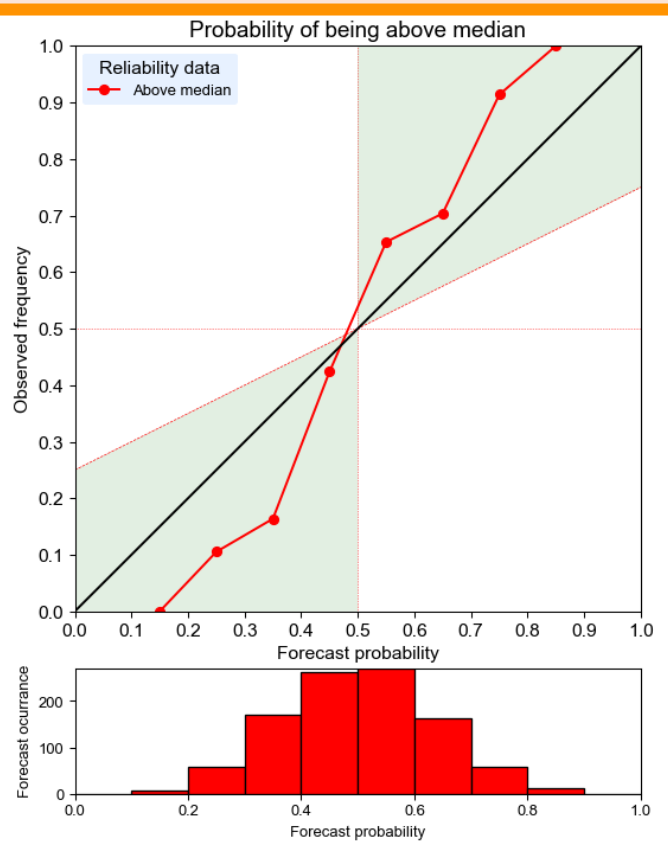
- Is the quantity (e.g. temperature T)
  above the median T_{med} of the historical observations?

Each forecast = a probability that $T > T_{med}$.
Over all observations, this is 50/50!

For a set of probability steps,
Pick all the times a forecast gave that probability,
And for those cases, count how often it actually occurred…

This is **underconfident**:

- Low probabilities aren't low enough
- High probabilities aren't high enough



Most forecasts have probabilities around the climatological frequency (50%)

Grant Agreement n. 776868

Reliability and probabilistic calibration

"Reliability" = how well the forecast probabilities match observed frequencies.

Can assess using a **reliability diagram**,
defined for the **particular event** we're trying to forecast.

This example:

- Is the quantity (e.g. temperature T)
  above the median T_{med} of the historical observations?

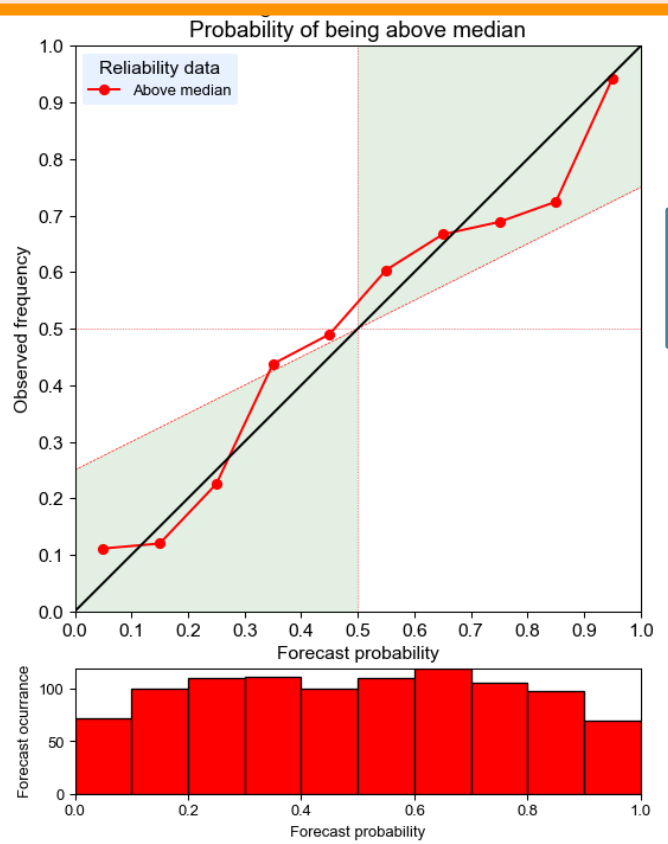
Each forecast = a probability that $T > T_{med}$
Over all observations, this is 50/50!

Difficulties:

- Can be very noisy – requires a lot of data
- Particular event of interest for users might not be clear.

This model is more **reliable**:

- probabilities close to observed frequencies
- Can be relied on for decision-making



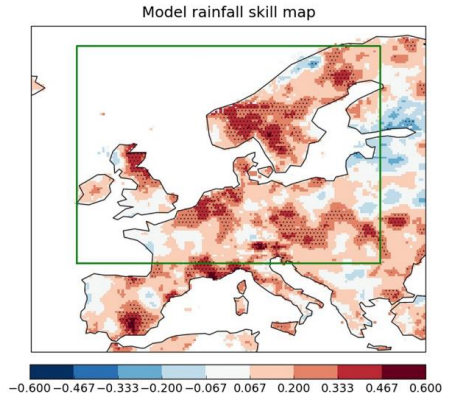
Model still not perfect:
Could calibrate probabilities further.

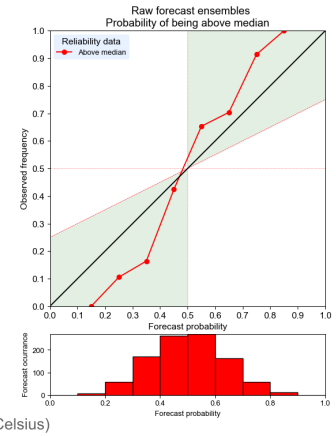
Forecast system can produce the whole range of probabilities

Grant Agreement n. 776868

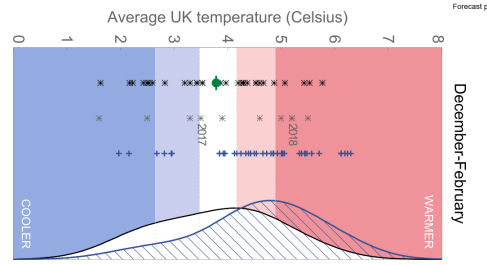
The Added Value of Seasonal Climate Forecasting for Integrated Risk Management

- Forecast uncertainty is usually ~ climatological variability
- Like the forecasts, skill is uncertain,
  estimated from limited historical forecasts + observations
- Expect to correct for mean bias, variance, …
  but also for **reliability** (probabilistic calibration)



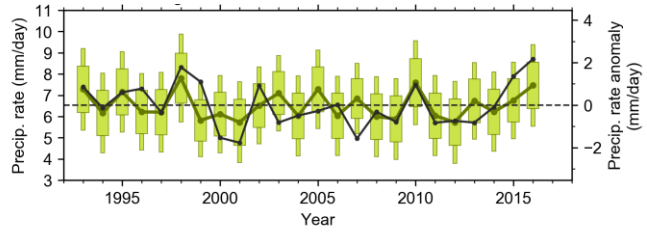


- Ideally, we'd just give forecasts of seasonal means averaged over large areas…
  But that's often not very useful! 😅



→ In **SECLI-FIRM** we are developing ways of using forecast model output, and relating it to user requirements to produce optimised **tailored forecasts**:

  - Teleconnections, downscaling, weather patterns, …



Grant Agreement n. 776868

Yes/No Events and Decision Making

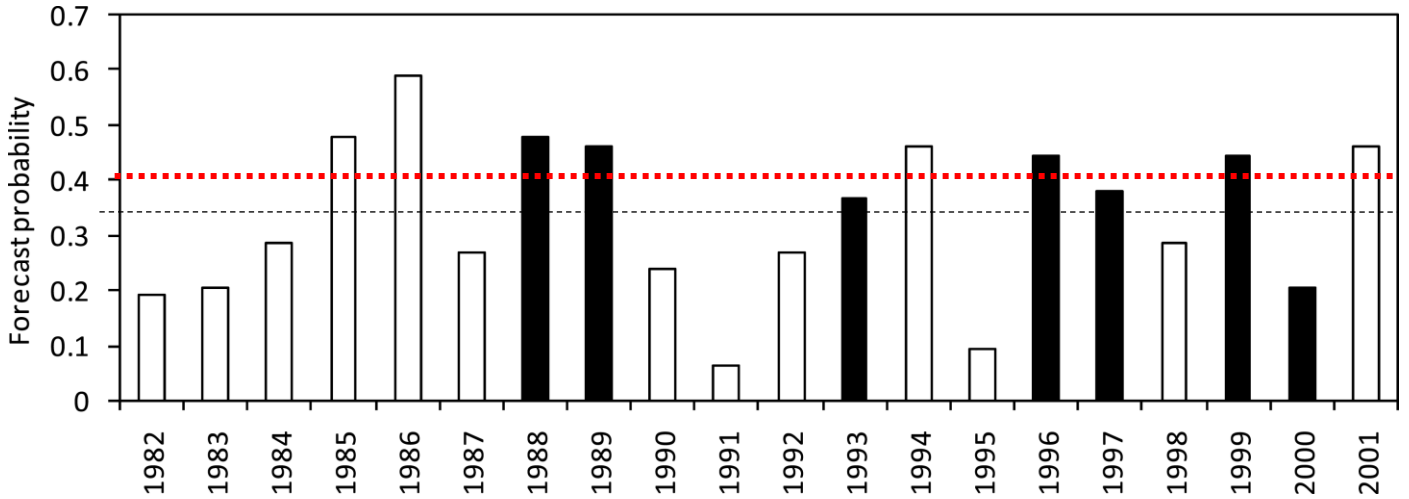
- Event is often a variable exceeding a threshold: $T > T_0$. → User defined, and observable
- Forecasts are probabilities of exceeding that threshold: $P(T>T_0)$
- Decision can be made given a probability threshold P_0, e.g. $P_0 = 70\%$
- Look for $P(T>T_0) \geq P_0$: If the chance of exceeding our T threshold is at least 70% then **take action**.

Observed $T<T_0$

Observed $T>T_0$

This example:
T_0 = upper tercile.
Climatological freq. = 33%.

Here: 7 cases / 20 (35%)



Example P_0 = 40%
Take action **8** times
No action **12** times
Hits: **4**
False Alarms: **4**
Misses: **3**
Correct rejections: **9**

HR = H /(H + M) = 4/7 = 57%
FAR= FA/(FA+CR) = 4/13 = 31%

Jones & Morse (2000), https://doi.org/10.1175/2010JCLI3208.1

Grant Agreement n. 776868