unseen_open

Timo Kelder

Sep 11, 2021

INTRODUCTION:

An open, reproducible and transferable workflow to assess and anticipate climate extremes beyond the observed record.

UNSEEN-open is an open source project using the *global* SEAS5 and ERA5 datasets. It makes evaluation of model simulations and extreme value analysis easy in order to **anticipate climate extremes** beyond the observed record. The project is developed as part of the ECMWF summer of weather code 2020 (esowc), which is funded by Copernicus.

UNSEEN-open relies on xarray for data preprocessing and uses ggplot and extRemes for the extreme value analysis. The extreme value utilities are being developed into an UNSEEN Rpackage.

CHAPTER

ONE

APPLICATIONS

In our recent NPJ Climate and Atmospheric Science paper we outline four potential applications where we believe UNSEEN might prove to be useful:

1. Help estimate design values, especially relevant for data scarce regions
2. Improve risk estimation of natural hazards by coupling UNSEEN to impact models
3. Detect trends in rare climate extremes
4. Increase our physical understanding of the drivers of (non-stationarity of) climate extremes

We hope this approach may see many applications across a range of scientific fields!

1.1 What is UNSEEN?

The UNprecedented Simulated Extreme ENsemble (UNSEEN, Thompson et al., 2017) approach is an increasingly popular method that exploits seasonal prediction systems to **assess and anticipate climate extremes beyond the observed record**. The approach uses pooled forecasts as plausible alternate realities. Instead of the 'single realization' of reality, pooled forecasts can be exploited to better assess the likelihood of infrequent events, which only have a limited chance of occurring in observed records. This method has for example been used to improve design levels of storm-surges in the river Rhine and to anticipate and understand heatwaves in China.

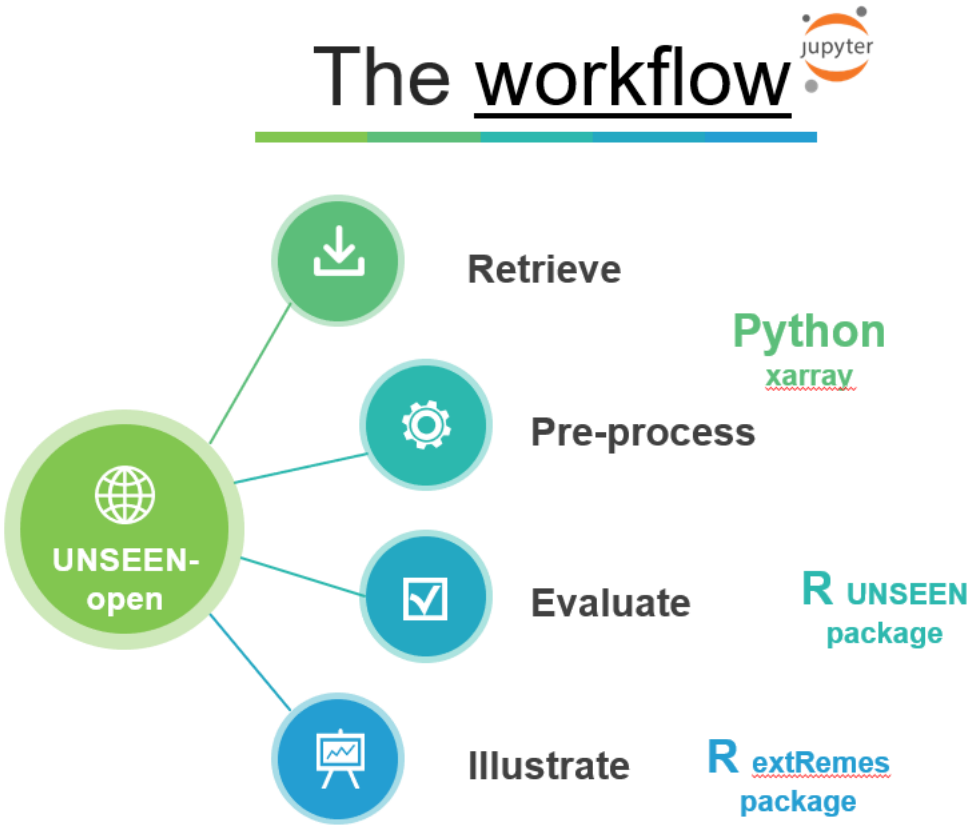
Recently, we applied this method to SEAS5 for the first time, with promising results for extreme precipitation over Norway and Svalbard (submitted to *NPJ Climate and Atmospheric Science*). SEAS5 is a good dataset for the UNSEEN approach because it is freely available and has a high resolution and a large ensemble compared to global climate models. To make this approach more accessible, we develop an open, reproducible and transferable workflow for the UNSEEN method using the open access SEAS5 dataset on the CDS.

1.2 UNSEEN-open

In this project, the aim is to build an open, reproducible, and transferable workflow for UNSEEN.

This means that **anyone** can assess **any** climate extreme event **anywhere** in the world!

The workflow consists of four steps, as illustrated below:



1.2.1 Overview

Here we provide an overview of the steps in UNSEEN-open.

Retrieve

We use *global open* Copernicus C3S data: the seasonal prediction system SEAS5 and the reanlysis ERA5.

The functions to retrieve all forecasts (SEAS5) and reanalysis (ERA5) are `retrieve_SEAS5` and `retrieve_ERA5`. You can select the climate variable, the target month(s) and the area - for more explanation see *retrieve*.

[2]:
```
retrieve.retrieve_SEAS5(
    variables=['2m_temperature', '2m_dewpoint_temperature'],
    target_months=[3, 4, 5],
    area=[70, -11, 30, 120],
    years=np.arange(1981, 2021),
    folder='../Siberia_example/SEAS5/')
```

[3]:
```
retrieve.retrieve_ERA5(variables=['2m_temperature', '2m_dewpoint_temperature'],
                       target_months=[3, 4, 5],
                       area=[70, -11, 30, 120],
                       folder='../Siberia_example/ERA5/')
```

Preprocess

In the preprocessing step, we first merge all downloaded files into one netcdf file. Then the rest of the preprocessing depends on the definition of the extreme event. For example, for the UK case study, we want to extract the UK average precipitation while for the Siberian heatwave we will just used the defined area to spatially average over. For the MAM season, we still need to take the seasonal average, while for the UK we already have the average February precipitation.

Read the docs on *preprocessing* for more info.

Evaluate

The evaluation step is important to assess whether the forecasts are realistic and consistent to the observations. There are three statistical tests available through the UNSEEN R package. See the *evaluation section* and also this paper for more info.

Illustrate

So what can we learn from UNSEEN-open?

In this section we apply extreme value theory to illustrate the applications. Have a look at the examples!

1.3 Examples

In this project, UNSEEN-open is applied to assess two extreme events in 2020: February 2020 UK precipitation and the 2020 Siberian heatwave.

Launch in Binder

1.3.1 Siberian Heatwave

The 2020 Siberian heatwave was a prolonged event that consistently broke monthly temperature the records. We show a gif of the temperature rank within the observations from 1979-2020, see *this section* for details. - Rank 1 mean highest on record - Rank 2 means second highest - etc..

This attribution study by World Weather Attribution (WWA) has shown that the event was made much more likely (600x) because of human induced climate change but also that the event was a very rare event within our present climate.

Could such an event be anticipated with UNSEEN?

With UNSEEN-open, we can assess whether extreme events like the Siberian heatwave have been forecasted already, i.e. whether we can anticipate such an event by exploiting all forecasts over the domain.

Retrieve data

The main functions to retrieve all forecasts (SEAS5) and reanalysis (ERA5) are `retrieve_SEAS5` and `retrieve_ERA5`. We want to download 2m temperature, for the March-May target months over the Siberian domain. By default, the hindcast years of 1981-2016 are downloaded for SEAS5. We include the years 1981-2020. The folder indicates where the files will be stored, in this case outside of the UNSEEN-open repository, in a 'Siberia_example' directory. For more explanation, see *retrieve*.

```
[ ]: retrieve.retrieve_SEAS5(
         variables=['2m_temperature', '2m_dewpoint_temperature'],
         target_months=[3, 4, 5],
         area=[70, -11, 30, 120],
         years=np.arange(1981, 2021),
         folder='../Siberia_example/SEAS5/')
```

```
[ ]: retrieve.retrieve_ERA5(variables=['2m_temperature', '2m_dewpoint_temperature'],
                           target_months=[3, 4, 5],
                           area=[70, -11, 30, 120],
                           folder='../Siberia_example/ERA5/')
```

Preprocess

In the preprocessing step, we first merge all downloaded files into one xarray dataset, then take the spatial average over the domain and a temporal average over the MAM season. Read the docs on *preprocessing* for more info.

```
[ ]: SEAS5_Siberia = preprocess.merge_SEAS5(folder = '../Siberia_example/SEAS5/', target_
     ↪months = [3,4,5])
```

And for ERA5:

```
[ ]: ERA5_Siberia = xr.open_mfdataset('../Siberia_example/ERA5/ERA5_????.nc',combine='by_
     ↪coords')
```

Then we calculate the day-in-month weighted seasonal average:

```
[ ]: SEAS5_Siberia_weighted = preprocess.season_mean(SEAS5_Siberia, years = 39)
     ERA5_Siberia_weighted = preprocess.season_mean(ERA5_Siberia, years = 42)
```

And we select the 2m temperature, and take the average over a further specified domain. This is a simple average, an area-weighed average is more appropriate, since grid cell area decreases with latitude, see *preprocess*.

```
[ ]: SEAS5_Siberia_events_zoomed = (
         SEAS5_Siberia_weighted['t2m'].sel(
             latitude=slice(70, 50),
             longitude=slice(65, 120)).
         mean(['longitude', 'latitude']))

     SEAS5_Siberia_events_zoomed_df = SEAS5_Siberia_events_zoomed.to_dataframe()
```

```
[ ]: ERA5_Siberia_events_zoomed = (
         ERA5_Siberia_weighted['t2m'].sel(  # Select 2 metre temperature
             latitude=slice(70, 50),        # Select the latitudes
             longitude=slice(65, 120)).     # Select the longitude
         mean(['longitude', 'latitude']))
```

(continues on next page)

(continued from previous page)

```
ERA5_Siberia_events_zoomed_df = ERA5_Siberia_events_zoomed.to_dataframe()
```

Evaluate

Note

From here onward we use R and not python!

We switch to R since we believe R has a better functionality in extreme value statistics.

Is the UNSEEN ensemble realistic?

To answer this question, we perform three statistical tests: independence, model stability and model fidelity tests. These statistical tests are available through the UNSEEN R package. See *evaluation* for more info.

[12]:
```
require(UNSEEN)
require(ggplot2)
```

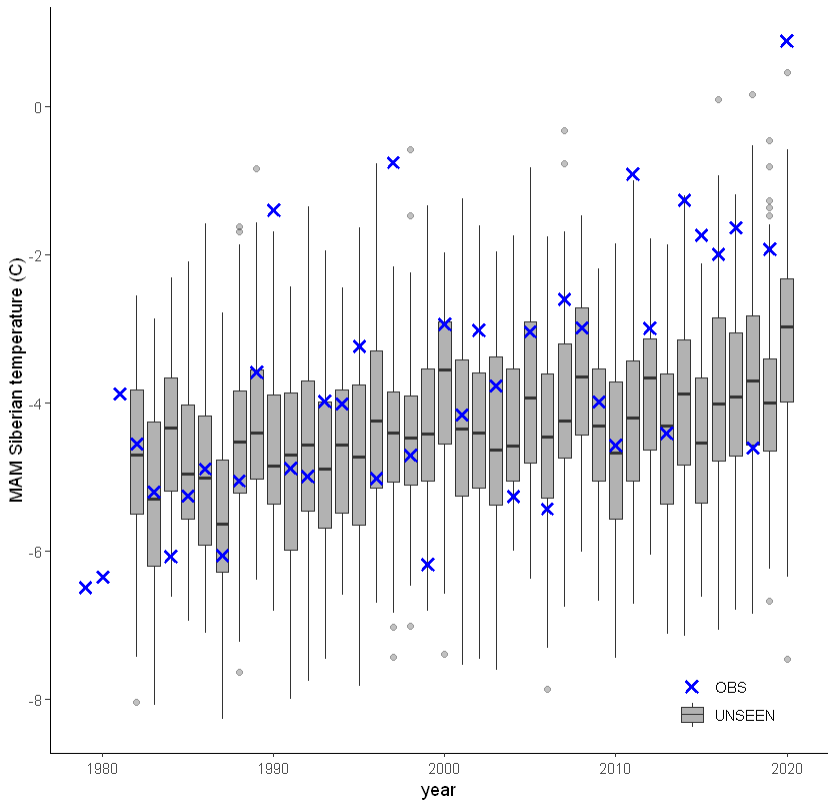
```
Loading required package: ggplot2
```

Timeseries

We plot the timeseries of SEAS5 (UNSEEN) and ERA5 (OBS) for the the Siberian Heatwave.

[13]:
```
timeseries = unseen_timeseries(
    ensemble = SEAS5_Siberia_events_zoomed_df,
    obs = ERA5_Siberia_events_zoomed,
    ensemble_yname = "t2m",
    ensemble_xname = "year",
    obs_yname = "t2m",
    obs_xname = "year",
    ylab = "MAM Siberian temperature (C)")

timeseries

# ggsave(timeseries, height = 5, width = 6,   filename = "graphs/Siberia_timeseries.
↪png")
```

```
Warning message:
"Removed 2756 rows containing non-finite values (stat_boxplot)."
Warning message:
"Removed 2756 rows containing non-finite values (stat_boxplot)."
```



The timeseries consist of **hindcast (years 1982-2016)** and **archived forecasts (years 2017-2020)**. The datasets are slightly different: the hindcasts contains 25 members whereas operational forecasts contain 51 members, the native resolution is different and the dataset from which the forecasts are initialized is different.

For the evaluation of the UNSEEN ensemble we want to only use the SEAS5 hindcasts for a consistent dataset. Note, 2017 is not used in either the hindcast nor the operational dataset, since it contains forecasts both initialized in 2016 (hindcast) and 2017 (forecast), see *retrieve*. We split SEAS5 into hindcast and operational forecasts:

[5]:
```
SEAS5_Siberia_events_zoomed_hindcast <- SEAS5_Siberia_events_zoomed_df[
    SEAS5_Siberia_events_zoomed_df$year < 2017 &
    SEAS5_Siberia_events_zoomed_df$number < 25,]

SEAS5_Siberia_events_zoomed_forecasts <- SEAS5_Siberia_events_zoomed_df[
    SEAS5_Siberia_events_zoomed_df$year > 2017,]
```

And we select the same years for ERA5.

[6]:
```
ERA5_Siberia_events_zoomed_hindcast <- ERA5_Siberia_events_zoomed[
    ERA5_Siberia_events_zoomed$year < 2017 &
    ERA5_Siberia_events_zoomed$year > 1981,]
```

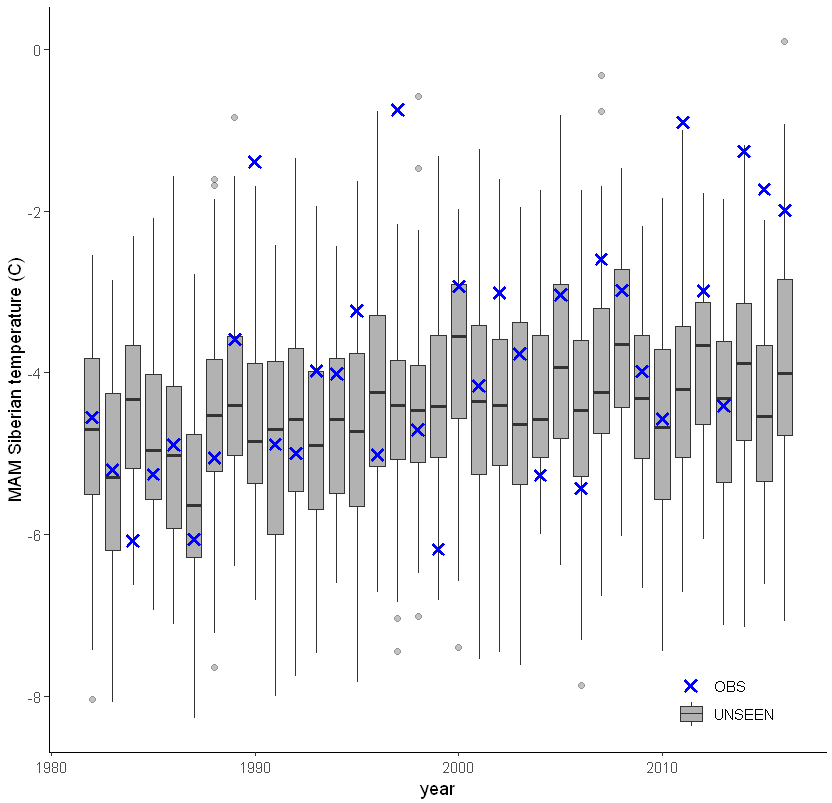
Which results in the following timeseries:

[7]:
```
unseen_timeseries(
    ensemble = SEAS5_Siberia_events_zoomed_hindcast,
    obs = ERA5_Siberia_events_zoomed_hindcast,
    ensemble_yname = "t2m",
    ensemble_xname = "year",
```

(continues on next page)

(continued from previous page)

```
    obs_yname = "t2m",
    obs_xname = "year",
    ylab = "MAM Siberian temperature (C)")
```



Evaluation tests

With the hindcast dataset we evaluate the independence, stability and fidelity. Here, we plot the results for the fidelity test, for more detail on the other tests see the *evaluation section*.

The fidelity test shows us how consistent the model simulations of UNSEEN (SEAS5) are with the observed (ERA5). The UNSEEN dataset is much larger than the observed – hence they cannot simply be compared. For example, what if we had faced a few more or a few less heatwaves purely by chance?

This would influence the observed mean, but not so much influence the UNSEEN ensemble because of the large data sample. Therefore we express the UNSEEN ensemble as a range of plausible means, for data samples of the same length as the observed. We do the same for higher order statistical moments.

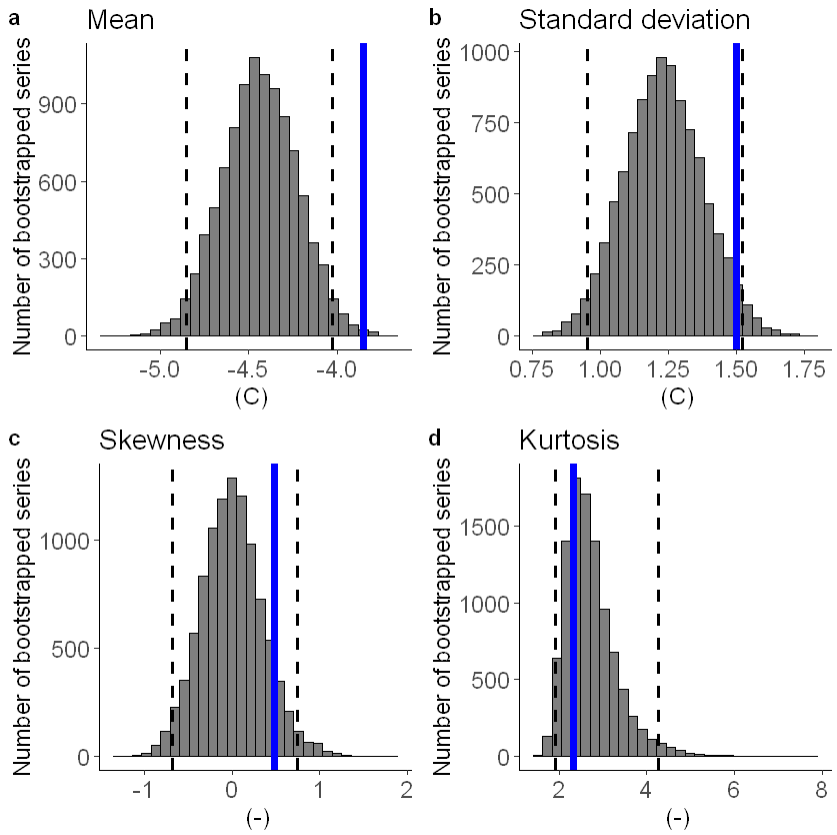
[15]:
```
Eval = fidelity_test(
    obs = ERA5_Siberia_events_zoomed_hindcast$t2m,
    ensemble = SEAS5_Siberia_events_zoomed_hindcast$t2m,
    units = 'C',
    biascor = FALSE,
    fontsize = 14
)
```

(continues on next page)

(continued from previous page)

```
Eval
ggsave(Eval, filename = "graphs/Siberia_fidelity.png")
```

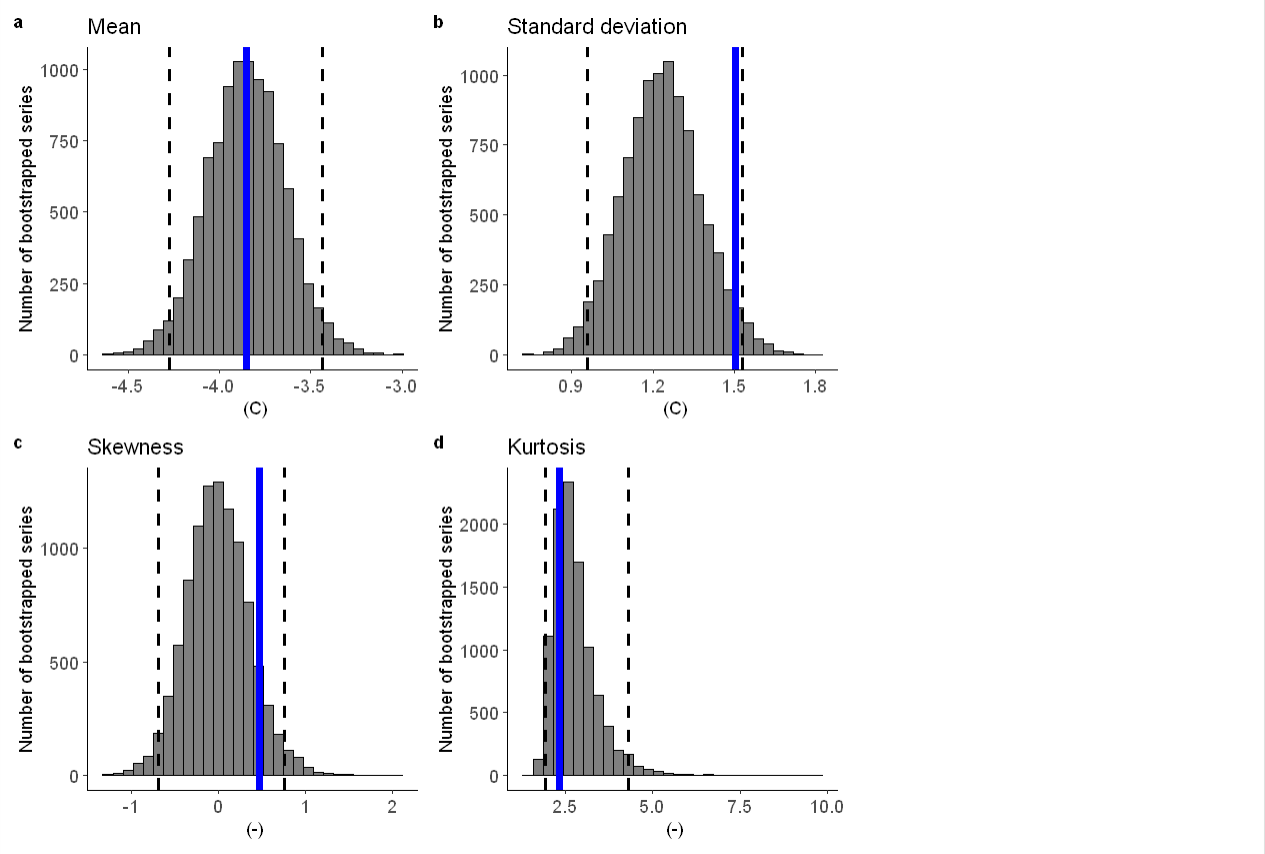
```
Saving 6.67 x 6.67 in image
```



The fidelity test shows that the mean of the UNSEEN ensemble is too low compared to the observed – the blue line falls outside of the model range in a. To correct for this low bias, we can apply an additive bias correction, which only corrects the mean of the simulations.

Lets apply the additive biascor:

[9]:
```
obs = ERA5_Siberia_events_zoomed_hindcast$t2m
ensemble = SEAS5_Siberia_events_zoomed_hindcast$t2m
ensemble_biascor = ensemble + (mean(obs) - mean(ensemble))

fidelity_test(
    obs = obs,
    ensemble = ensemble_biascor,
    units = 'C',
    biascor = FALSE
)
```



This shows us what we expected: the mean bias is corrected because the model simulations are shifted up (the blue line is still the same, the axis has just shifted along with the histogram), but the other statistical moments are the same.

Illustrate

[10]:
```
source('src/evt_plot.r')
```

```
Loading required package: Lmoments

Loading required package: distillery


Attaching package: 'extRemes'


The following objects are masked from 'package:stats':

    qqnorm, qqplot
```

First, we fit a Gumbel and a GEV distribution (including shape parameter) to the observed extremes. The Gumbel distribution best describes the data because the p-value of 0.65 is much above 0.05 (based on the likelihood ratio test).

[11]:
```
fit_obs_Gumbel <- fevd(x = ERA5_Siberia_events_zoomed_hindcast$t2m,
                       type = "Gumbel"
```

(continues on next page)

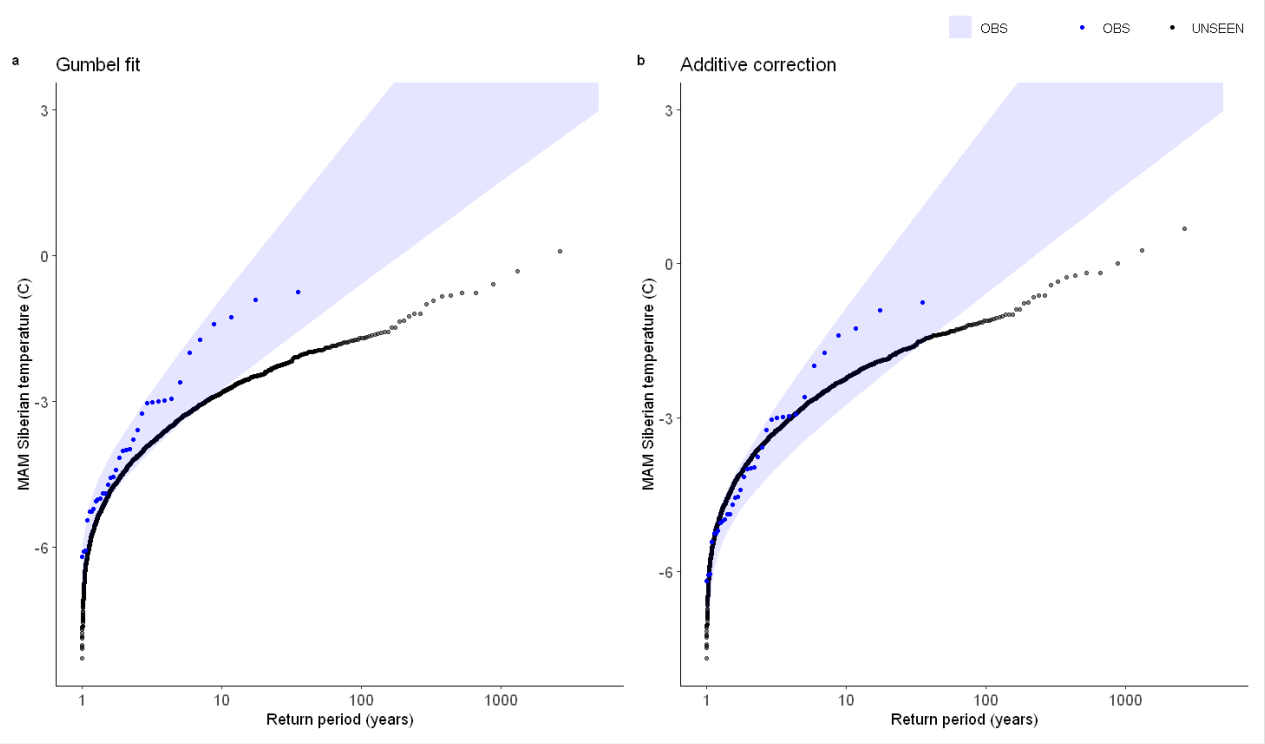
(continued from previous page)

```
                    )
fit_obs_GEV <- fevd(x = ERA5_Siberia_events_zoomed_hindcast$t2m,
                    type = "GEV"
                    )
lr.test(fit_obs_Gumbel, fit_obs_GEV)
```

```
        Likelihood-ratio Test

data:  ERA5_Siberia_events_zoomed_hindcast$t2mERA5_Siberia_events_zoomed_hindcast$t2m
Likelihood-ratio = 0.21004, chi-square critical value = 3.8415, alpha =
0.0500, Degrees of Freedom = 1.0000, p-value = 0.6467
alternative hypothesis: greater
```

We show the gumbel plot for the observed (ERA5) and UNSEEN (SEAS5 hindcast data). This shows that the UNSEEN simulations are not within the uncertainty range of the observations. This has likely two reasons, illustrated in the evaluation section: there is some dependence between the events and there is too little variability within the UNSEEN ensemble.

[12]:
```
options(repr.plot.width = 12)
GEV_hindcast <- EVT_plot(ensemble = SEAS5_Siberia_events_zoomed_hindcast$t2m,
                         obs = ERA5_Siberia_events_zoomed_hindcast$t2m,
                         main = "Gumbel fit",
                         GEV_type = "Gumbel",
                         ylim = 3,
                         y_lab = 'MAM Siberian temperature (C)'
                        )
GEV_hindcast_corrected <- EVT_plot(ensemble = ensemble_biascor, #SEAS5_Siberia_events_
↪zoomed_hindcast$t2m,
                                   obs = ERA5_Siberia_events_zoomed_hindcast$t2m,
                                   main = "Additive correction",
                                   GEV_type = "Gumbel",
                                   ylim = 3,
                                   y_lab = 'MAM Siberian temperature (C)'
                                  )

ggarrange(GEV_hindcast, GEV_hindcast_corrected,
  labels = c("a", "b"), # ,"c","d"),
  common.legend = T,
  font.label = list(size = 10, color = "black", face = "bold", family = NULL),
  ncol = 2, nrow = 1
)
```

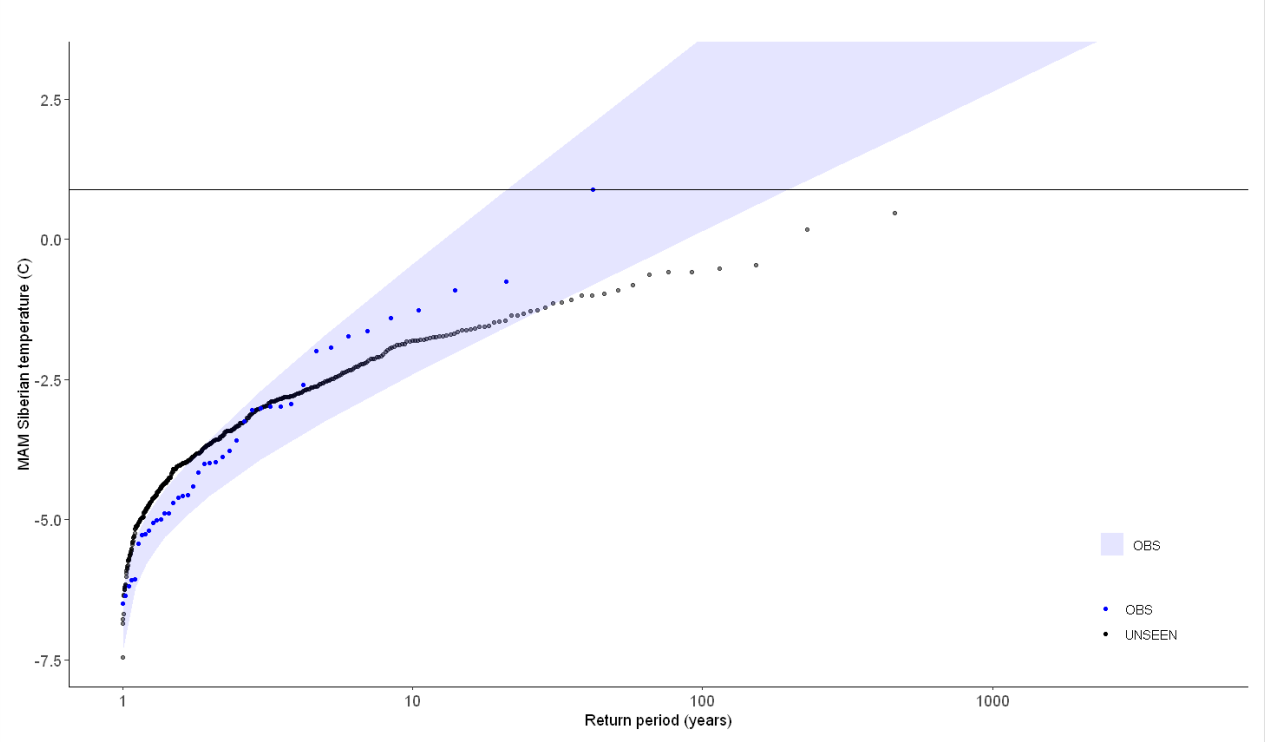


So what can we get out of it? What if we look at the operational forecast? Even if we cannot use the dataset as a whole to estimate the likelihood of occurrence, have events similar to the 2020 event occurred?

We select 2018-2020 archived SEAS5 forecasts as UNSEEN events, check out the *timeseries section* for more info on the difference between the hindcast data and the archived forecasts. We furthermore select all ERA5 (observed) events except for the 2020 event as reference to estimate the likelihood of the 2020 event.

[13]:
```
ERA5_Siberia_events_zoomed_min1 <- ERA5_Siberia_events_zoomed[1:length(ERA5_Siberia_
↪events_zoomed$t2m)-1,]
ERA5_Siberia_events_zoomed_2020 <- ERA5_Siberia_events_zoomed[length(ERA5_Siberia_
↪events_zoomed$t2m),]
```

[14]:
```
GEV_forecasts <- EVT_plot(ensemble = SEAS5_Siberia_events_zoomed_forecasts$t2m,
                          obs = ERA5_Siberia_events_zoomed$t2m,
                          main = "",
                          GEV_type = "Gumbel",
                          ylim = 3,
                          y_lab = 'MAM Siberian temperature (C)'
                         ) # %>%
GEV_forecasts + geom_hline(yintercept = ERA5_Siberia_events_zoomed_2020$t2m)
```



Applications:

Prolonged heat events with an average temperature above 0 degrees over Siberia can have enormous impacts on the local environment, such as wildfires, invasion of pests and infrastructure failure, and on the global environment, through the release of greenhouse gasses during permafrost thawing.

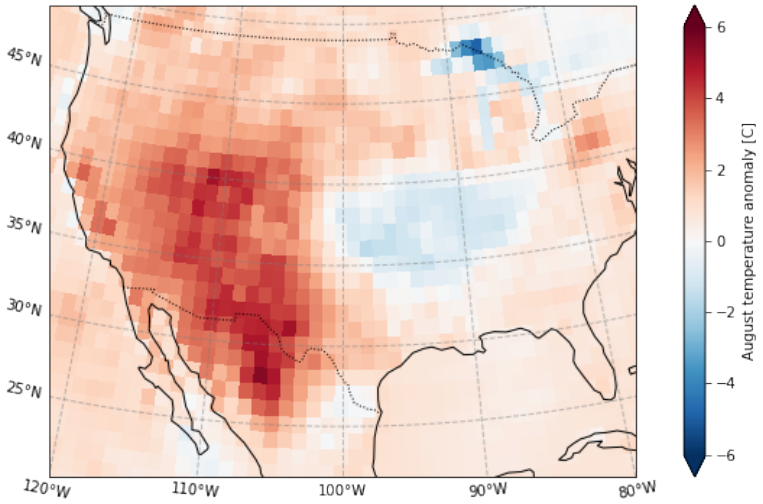
With UNSEEN-open, we can: 1. *Assess the drivers* of the most severe events. The 2020 event seemed to be caused by a very anomalous Indian Ocean Dipole (IOD). What can we learn from the highest UNSEEN events? To what extent are these also driven by an anomalous IOD or are there other drivers of such severe heat events? 2. Perform *nonstationary analysis* in rare extremes, such as the 100-year event. There seems to be a trend over the hindcast period in the severe heatwaves. We can perform non-stationary analysis to estimate the change in the magnitude and frequency of the heatwaves and, if we find a change, we could explore the drivers of this change. 3. *Evaluate forecasts*. Since we are using seasonal forecasts in this setup, we could explore the forecast skill in simulating heatwaves over Siberia.

Launch in Binder

1.3.2 California fires

In August 2020 in California, wildfires have burned more than a million acres of land. This years' fire season was also unique in the number of houses destroyed.

Here we retrieve average august temperatures over California within ERA5 1979-2020 and show anomalous August



2020 was.

We furthermore create an UNSEEN ensemble and show that these kind of fire seasons can be expected to occur more often in our present climate since we find a clear trend in temperature extremes over the last decades.

Retrieve data

The main functions to retrieve all forecasts (SEAS5) and reanalysis (ERA5) are `retrieve_SEAS5` and `retrieve_ERA5`. We want to download 2m temperature for August over California. By default, the hindcast years of 1981-2016 are downloaded for SEAS5. We include the years 1981-2020. The folder indicates where the files will be stored, in this case outside of the UNSEEN-open repository, in a 'California_example' directory. For more explanation, see *retrieve*.

[1]:
```
import os
import sys
sys.path.insert(0, os.path.abspath('../../../'))
os.chdir(os.path.abspath('../../../'))

import src.cdsretrieve as retrieve
import src.preprocess as preprocess

import numpy as np
import xarray as xr
```

[2]:
```
retrieve.retrieve_SEAS5(
    variables=['2m_temperature', '2m_dewpoint_temperature'],
    target_months=[8],
    area=[70, -130, 20, -70],
    years=np.arange(1981, 2021),
    folder='E:/PhD/California_example/SEAS5/')
```

[3]:
```
retrieve.retrieve_ERA5(variables=['2m_temperature', '2m_dewpoint_temperature'],
                       target_months=[8],
                       area=[70, -130, 20, -70],
                       folder='E:/PhD/California_example/ERA5/')
```

Preprocess

In the preprocessing step, we first merge all downloaded files into one xarray dataset, then take the spatial average over the domain and a temporal average over the MAM season. Read the docs on *preprocessing* for more info.

[4]:
```
SEAS5_California = preprocess.merge_SEAS5(folder ='E:/PhD/California_example/SEAS5/',␣
↪target_months = [8])
```

```
Lead time: 07
6
5
4
3
```

And for ERA5:

[5]:
```
ERA5_California = xr.open_mfdataset('E:/PhD/California_example/ERA5/ERA5_????.nc',
↪combine='by_coords')
```

We calculate the *standardized anomaly of the 2020 event* and select the 2m temperature over the region where 2 standard deviations from the 1979-2010 average was exceeded. This is a simple average, an area-weighed average is more appropriate, since grid cell area decreases with latitude, see *preprocess*.

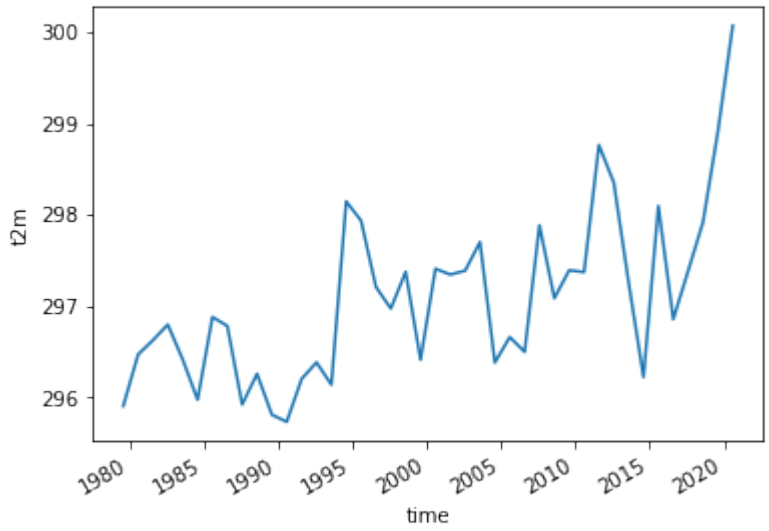
[6]:
```
ERA5_anomaly = ERA5_California['t2m'] - ERA5_California['t2m'].sel(time=slice('1979',
↪'2010')).mean('time')
ERA5_sd_anomaly = ERA5_anomaly / ERA5_California['t2m'].std('time')
```

[7]:
```
ERA5_California_events = (
    ERA5_California['t2m'].sel(  # Select 2 metre temperature
        longitude = slice(-125,-100),     # Select the longitude
        latitude = slice(45,20)).         # And the latitude
    where(ERA5_sd_anomaly.sel(time = '2020').squeeze('time') > 2). ##Mask the region␣
↪where 2020 sd >2.
    mean(['longitude', 'latitude'])) #And take the mean
```

Plot the August temperatures over the defined California domain:

[8]:
```
ERA5_California_events.plot()
```

[8]:
```
[<matplotlib.lines.Line2D at 0x1dc77e84910>]
```



Select the same domain for SEAS5 and extract the events.

[9]:
```
SEAS5_California_events = (
    SEAS5_California['t2m'].sel(
        longitude = slice(-125,-100),     # Select the longitude
        latitude = slice(45,20)).
    where(ERA5_sd_anomaly.sel(time = '2020').squeeze('time') > 2).
    mean(['longitude', 'latitude']))
```

And here we store the data in the Data section so the rest of the analysis in R can be reproduced.

[10]:
```
SEAS5_California_events.to_dataframe().to_csv('Data/SEAS5_California_events.csv')
ERA5_California_events.to_dataframe().to_csv('Data/ERA5_California_events.csv')
```

```
C:\anaconda3\envs\unseen\lib\site-packages\dask\array\numpy_compat.py:41:␣
↪RuntimeWarning: invalid value encountered in true_divide
  x = np.divide(x1, x2, out)
```

Evaluate

Note

From here onward we use R and not python!

We switch to R since we believe R has a better functionality in extreme value statistics.

Is the UNSEEN ensemble realistic?

To answer this question, we perform three statistical tests: independence, model stability and model fidelity tests. These statistical tests are available through the UNSEEN R package. See *evaluation* for more info.

[2]:
```
require(UNSEEN)
```

```
Loading required package: UNSEEN
```

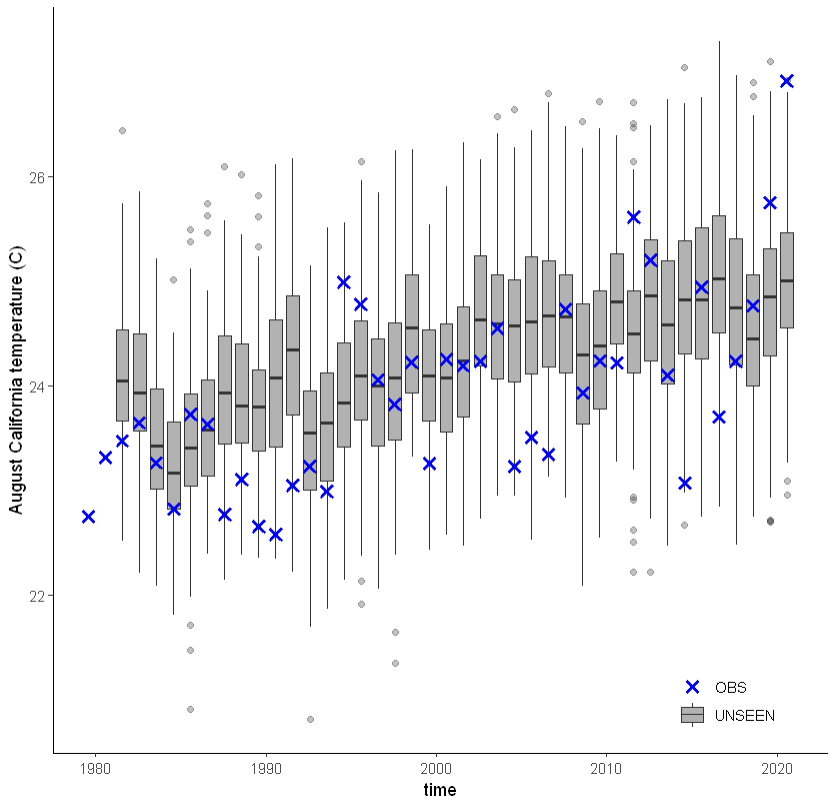
Timeseries

We plot the timeseries of SEAS5 (UNSEEN) and ERA5 (OBS) for the the Siberian Heatwave.

[3]:
```
timeseries = unseen_timeseries(
    ensemble = SEAS5_California_events,
    obs = ERA5_California_events,
    ensemble_yname = "t2m",
    ensemble_xname = "time",
    obs_yname = "t2m",
    obs_xname = "time",
    ylab = "August California temperature (C)")


timeseries

ggsave(timeseries, height = 5, width = 6,   filename = "graphs/Calif_timeseries.png")
```

```
Warning message:
"Removed 4680 rows containing non-finite values (stat_boxplot)."
```

```
Error in ggsave(timeseries, height = 5, width = 6, filename = "graphs/Calif_
↪timeseries.png"): could not find function "ggsave"
Traceback:
```



The timeseries consist of **hindcast (years 1982-2016)** and **archived forecasts (years 2017-2020)**. The datasets are slightly different: the hindcasts contains 25 members whereas operational forecasts contain 51 members, the native resolution is different and the dataset from which the forecasts are initialized is different.

For the evaluation of the UNSEEN ensemble we want to only use the SEAS5 hindcasts for a consistent dataset. Note, 2017 is not used in either the hindcast nor the operational dataset, since it contains forecasts both initialized in 2016 (hindcast) and 2017 (forecast), see *retrieve*. We split SEAS5 into hindcast and operational forecasts:

[4]:
```
SEAS5_California_events_hindcast <- SEAS5_California_events[
    SEAS5_California_events$time < '2017-02-01' &
    SEAS5_California_events$number < 25,]

SEAS5_California_events_forecasts <- SEAS5_California_events[
    SEAS5_California_events$time > '2017-02-01',]
```

And we select the same years for ERA5.

[5]:
```
ERA5_California_events_hindcast <- ERA5_California_events[
    ERA5_California_events$time > '1981-02-01' &
    ERA5_California_events$time < '2017-02-01',]
```

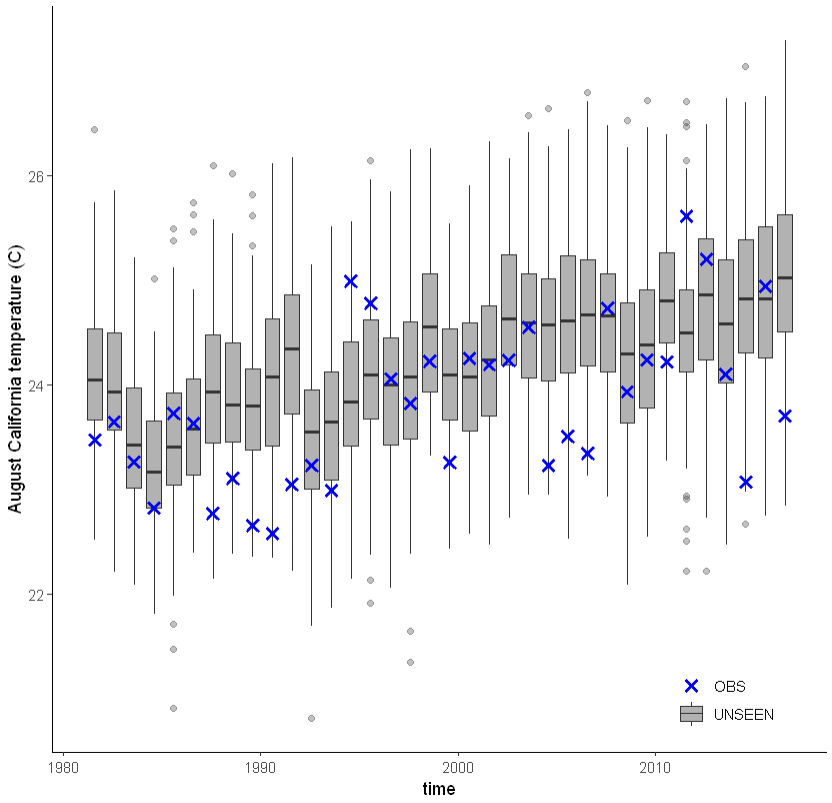
Which results in the following timeseries:

[6]:
```
unseen_timeseries(
    ensemble = SEAS5_California_events_hindcast,
    obs = ERA5_California_events_hindcast,
    ensemble_yname = "t2m",
    ensemble_xname = "time",
```

(continues on next page)

(continued from previous page)

```
    obs_yname = "t2m",
    obs_xname = "time",
    ylab = "August California temperature (C)")
```



Evaluation tests

With the hindcast dataset we evaluate the independence, stability and fidelity. Here, we plot the results for the fidelity test, for more detail on the other tests see the *evaluation section*.

The fidelity test shows us how consistent the model simulations of UNSEEN (SEAS5) are with the observed (ERA5). The UNSEEN dataset is much larger than the observed – hence they cannot simply be compared. For example, what if we had faced a few more or a few less heatwaves purely by chance?

This would influence the observed mean, but not so much influence the UNSEEN ensemble because of the large data sample. Therefore we express the UNSEEN ensemble as a range of plausible means, for data samples of the same length as the observed. We do the same for higher order statistical moments.

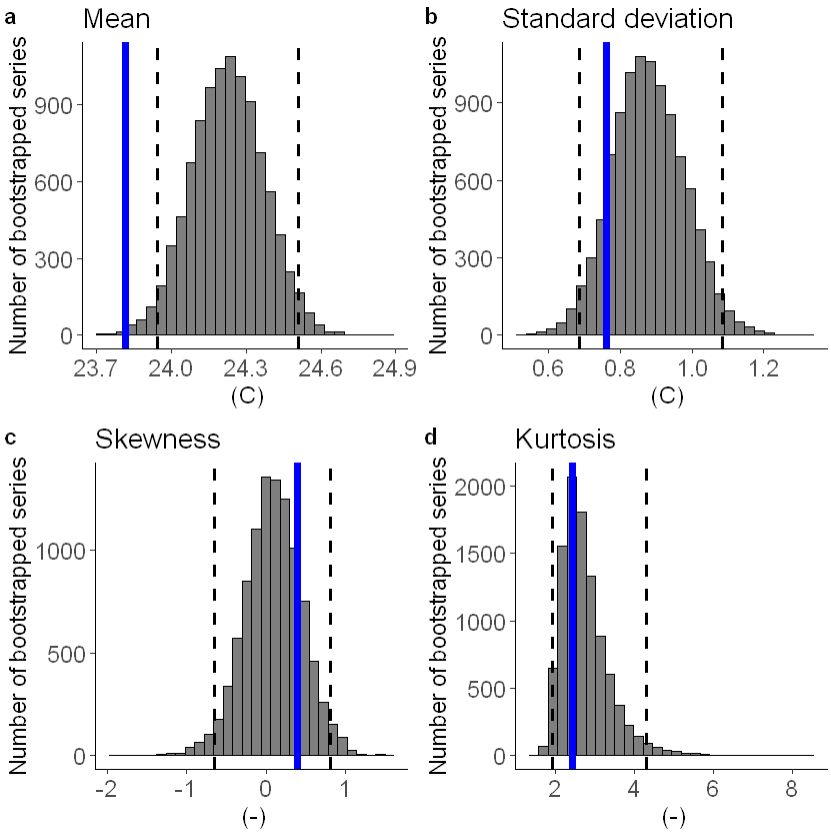
[7]:
```
Eval = fidelity_test(
    obs = ERA5_California_events_hindcast$t2m,
    ensemble = SEAS5_California_events_hindcast$t2m,
    units = 'C',
    biascor = FALSE,
    fontsize = 14
)
```

(continues on next page)

(continued from previous page)

```
Eval
ggsave(Eval, filename = "graphs/Calif_fidelity.png")
```

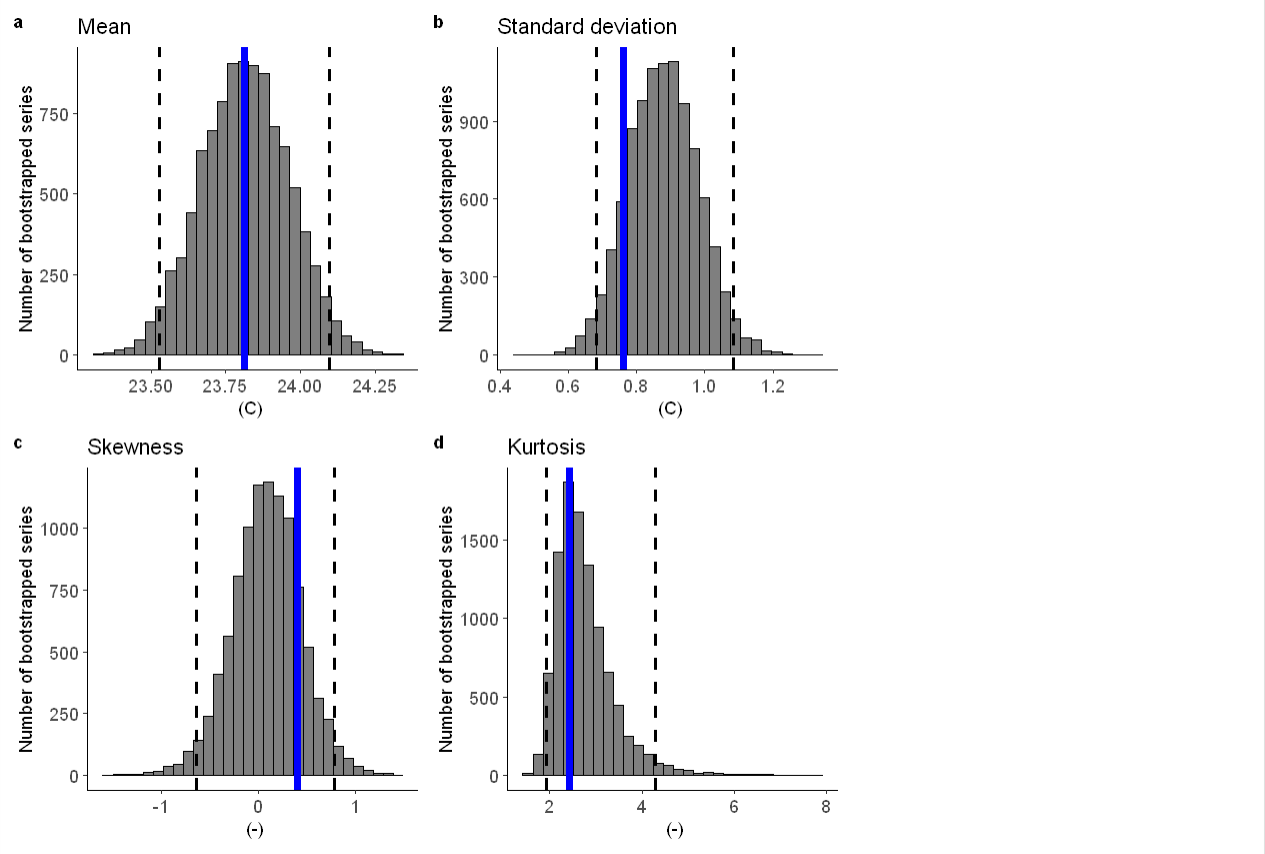
```
Error in ggsave(Eval, filename = "graphs/Calif_fidelity.png"): could not find␣
↪function "ggsave"
Traceback:
```



The fidelity test shows that the mean of the UNSEEN ensemble is too low compared to the observed – the blue line falls outside of the model range in a. To correct for this low bias, we can apply an additive bias correction, which only corrects the mean of the simulations.

Lets apply the additive biascor:

[10]:
```
obs = ERA5_California_events_hindcast$t2m
ensemble = SEAS5_California_events_hindcast$t2m
ensemble_biascor = ensemble + (mean(obs) - mean(ensemble))

fidelity_test(
    obs = obs,
    ensemble = ensemble_biascor,
    units = 'C',
    biascor = FALSE
)
```



This shows us what we expected: the mean bias is corrected because the model simulations are shifted up (the blue line is still the same, the axis has just shifted along with the histogram), but the other statistical moments are the same.

Illustrate

[8]:
```
source('src/evt_plot.r')
```

```
Loading required package: Lmoments

Loading required package: distillery


Attaching package: 'extRemes'


The following objects are masked from 'package:stats':

    qqnorm, qqplot
```

We apply extreme value theory to analyze the trend in 100-year temperature extremes. There are different extreme value distributions that can be used to fit to the data. First, we fit a stationary Gumbel and a GEV distribution (including shape parameter) to the observed extremes. Then we fit a nonstationary GEV distribution to the observed temperatures and show that this better describes the data because the p-value of 0.006 and 0.002 are very small (much below 0.05 based on 5% significance with the likelihood ratio test).

[11]:
```
## Fit stationary distributions
fit_obs_Gumbel <- fevd(x = obs,
                      type = "Gumbel"
                     )
fit_obs_GEV <- fevd(x = obs,
                    type = "GEV"
                   )
## And the nonstationary distribution
fit_obs_GEV_nonstat <- fevd(x = obs,
                            type = "GEV",
                            location.fun = ~ c(1:36), ##Fitting the gev with a␣
↪location and scale parameter linearly correlated to the covariate (years)
                            scale.fun = ~ c(1:36),
                            use.phi = TRUE
                           )
#And test the fit
##1. Stationary Gumbel vs stationary GEV
lr.test(fit_obs_Gumbel, fit_obs_GEV_nonstat)
##2. Stationary GEV vs Nonstationary GEV
lr.test(fit_obs_GEV, fit_obs_GEV_nonstat)
```

```
        Likelihood-ratio Test

data:  obsobs
Likelihood-ratio = 12.617, chi-square critical value = 7.8147, alpha =
0.0500, Degrees of Freedom = 3.0000, p-value = 0.005542
alternative hypothesis: greater
```

```
        Likelihood-ratio Test

data:  obsobs
Likelihood-ratio = 12.013, chi-square critical value = 5.9915, alpha =
0.0500, Degrees of Freedom = 2.0000, p-value = 0.002463
alternative hypothesis: greater
```

For the unseen ensemble this analysis is slightly more complicated since we need a covariate that has the same length as the ensemble:

[12]:
```
#Create the ensemble covariate
year_vector = as.integer(format(SEAS5_California_events_hindcast$time, format="%Y"))
covariate_ens = year_vector - 1980

# Fit the stationary distribution
fit_unseen_GEV <- fevd(x = ensemble_biascor,
                       type = 'GEV',
                       use.phi = TRUE)

fit_unseen_Gumbel <- fevd(x = ensemble_biascor,
                          type = 'Gumbel',
                          use.phi = TRUE)

# Fit the nonstationary distribution
fit_unseen_GEV_nonstat <- fevd(x = ensemble_biascor,
                               type = 'GEV',
```

(continues on next page)

(continued from previous page)

```
                                location.fun = ~ covariate_ens, ##Fitting the gev with␣
↪a location and scale parameter linearly correlated to the covariate (years)
                                scale.fun = ~ covariate_ens,
                                use.phi = TRUE)
```

And the likelihood ratio test tells us that the nonstationary GEV distribution is the best fit, both p-values < 2.2e-16:

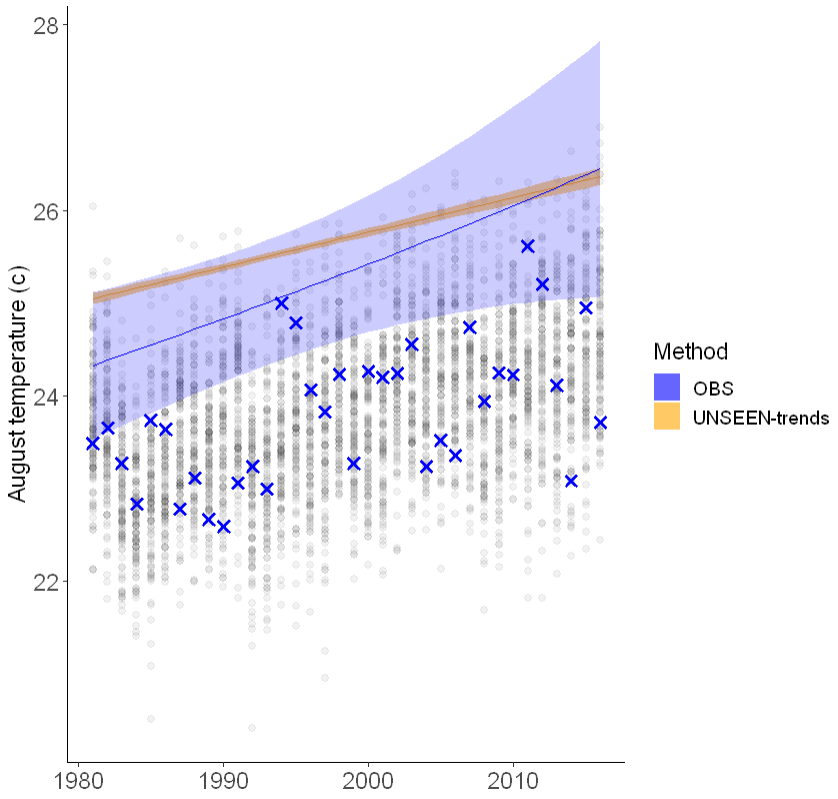
[13]:
```
#And test the fit
##1. Stationary Gumbel vs stationary GEV
lr.test(fit_unseen_Gumbel,fit_unseen_GEV)
##2. Stationary GEV vs Nonstationary GEV
lr.test(fit_unseen_GEV, fit_unseen_GEV_nonstat)
```

```
        Likelihood-ratio Test

data:  ensemble_biascorensemble_biascor
Likelihood-ratio = 577.13, chi-square critical value = 3.8415, alpha =
0.0500, Degrees of Freedom = 1.0000, p-value < 2.2e-16
alternative hypothesis: greater
```

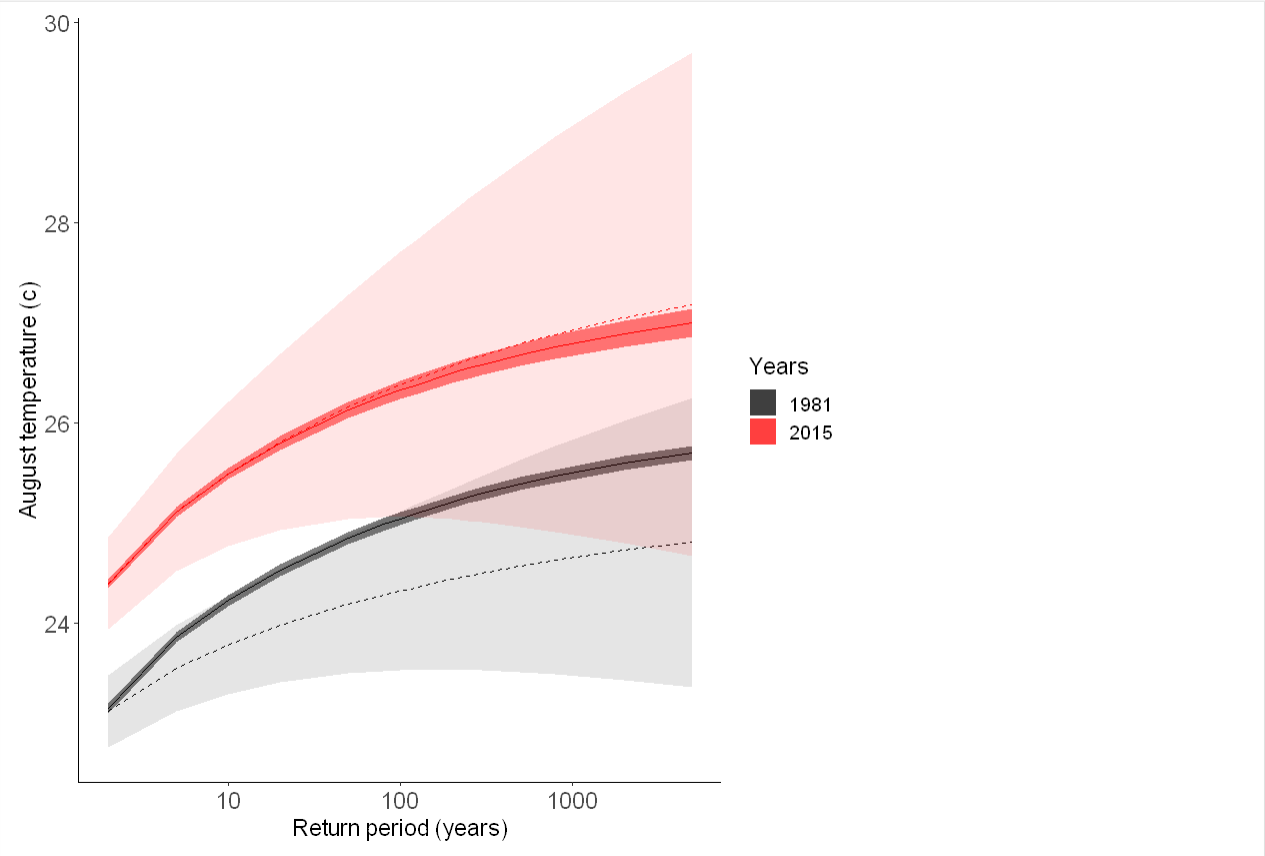
```
        Likelihood-ratio Test

data:  ensemble_biascorensemble_biascor
Likelihood-ratio = 956.98, chi-square critical value = 5.9915, alpha =
0.0500, Degrees of Freedom = 2.0000, p-value < 2.2e-16
alternative hypothesis: greater
```

We plot unseen trends in 100-year extremes. For more info on the methods see this paper

[14]:
```
p1 <- unseen_trends1(ensemble = ensemble_biascor,
                     x_ens = year_vector,
                     x_obs = 1981:2016,
                     rp = 100,
                     obs = obs,
                     covariate_ens = covariate_ens,
                     covariate_obs = c(1:36),
                     GEV_type = 'GEV',
                     ylab = 'August temperature (c)')
p1
```



[15]:
```
p2 <- unseen_trends2(ensemble = ensemble_biascor,
                     obs = obs,
                     covariate_ens = covariate_ens,
                     covariate_obs = c(1:36),
                     GEV_type = 'GEV',
                     ylab = 'August temperature (c)')

p2
```



Applications:

We have seen the worst fire season over California this year. Such fires are likely part of a chain of impacts, from droughts to heatwaves to fires, with feedbacks between them. Here we assess August temperatures and show that the 2020 August average temperature was very anomalous. We furthermore use SEAS5 forecasts to analyze the trend in rare extremes. Evaluation metrics show that the model simulations have a high bias, which we correct for using an additive bias correction. UNSEEN trend analysis shows a clear trend over time, both in the model and in the observed temperatures. Based on this analysis, temperature extremes that you would expect to occur once in 100 years in 1981 might occur once in 10 years in 2015 – and even more frequently now!

Note

Our analysis shows the results of a *linear* trend analysis of August temperature averages over 1981-2015. Other time windows, different trends than linear, and spatial domains could (should?) be investigated, as well as drought estimates in addition to temperature extremes.

Launch in Binder

1.3.3 UK Precipitation

February 2020 case study

February 2020 was the wettest February on record in the UK (since 1862), according to the Met Office. The UK faced three official storms during February, and this exceptional phenomena attracted media attention, such as an article from the BBC on increased climate concerns among the population. A Carbon Brief post explained why the UK saw such record-breaking rainfall and put this rare event into perspective, citing, amongst other approaches, the UNSEEN method. The UNSEEN study by Thompson et al., 2017 assessed monthly precipitation over the UK. They showed that the monthly precipitation records for south east England have a 7% chance of being exceeded in at least one month in any given winter. They did not use SEAS5 but the Met Office model ensemble. This work was taken up in the National Flood Resilience Review (2016), showing the high relevance and applicability of the method.

Here, the aim is to **build an open, reproducible and transferable workflow**, that will be tested for this well-studied region of the world and can be transferred to other regions and climate variables of interest, such as the 2020 Siberian heat and California fires.

Retrieve data

The main functions to retrieve all forecasts (SEAS5) is `retrieve_SEAS5`. We want to download February average precipitation over the UK. By default, the hindcast years of 1981-2016 are downloaded for SEAS5. The folder indicates where the files will be stored, in this case outside of the UNSEEN-open repository, in a 'UK_example' directory. For more explanation, see retrieve.

[ ]:
```
retrieve.retrieve_SEAS5(variables = 'total_precipitation',
                        target_months = [2],
                        area = [60, -11, 50, 2],
                        folder = '../UK_example/SEAS5/')
```

We use the EOBS observational dataset to evaluate the UNSEEN ensemble. I tried to download EOBS through the Copernicus Climate Data Store, but the Product is temporally disabled for maintenance purposes. As workaround I downloaded EOBS (from 1950 - 2019) and the most recent EOBS data (2020) here. Note, you have to register as E-OBS user.

Preprocess

In the preprocessing step, we first merge all downloaded files into one xarray dataset, see preprocessing.

[ ]:
```
SEAS5 = xr.open_dataset('../UK_example/SEAS5/SEAS5.nc')
```

I tried to download EOBS through CDS, but the Product was temporally disabled for maintenance purposes (1.2 Retrieve). As workaround, here, I downloaded EOBS (from 1950 - 2019) and the most recent EOBS data (2020) here. Note, you have to register as E-OBS user.

I will select February monthly mean precipitation to compare to SEAS5. I have taken the average mm/day over the month, which I think is more fair than the total monthly precipitation because of leap days.

[5]:
```
EOBS = xr.open_dataset('../UK_example/EOBS/rr_ens_mean_0.25deg_reg_v20.0e.nc') ##␣
↪open the data
EOBS = EOBS.resample(time='1m').mean() ## Monthly averages
EOBS = EOBS.sel(time=EOBS['time.month'] == 2) ## Select only February
EOBS
```

```
/soge-home/users/cenv0732/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/xarray/
↪core/nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

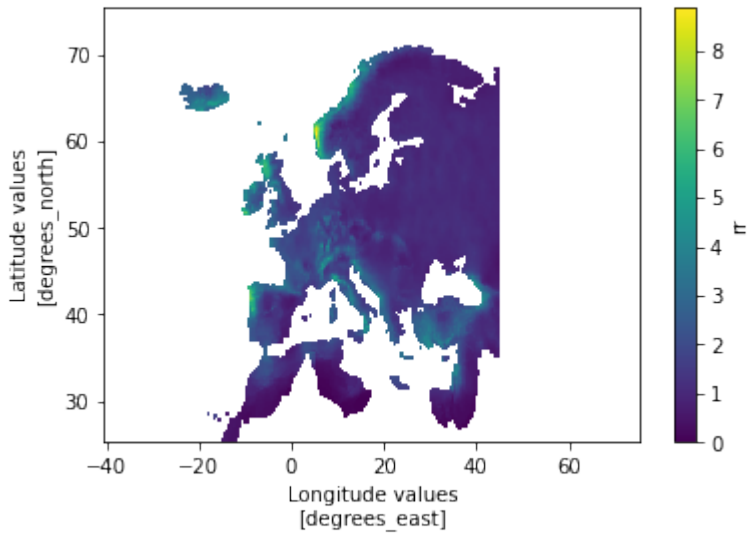
[5]:
```
<xarray.Dataset>
Dimensions:    (latitude: 201, longitude: 464, time: 70)
Coordinates:
  * time       (time) datetime64[ns] 1950-02-28 1951-02-28 ... 2019-02-28
  * longitude  (longitude) float64 -40.38 -40.12 -39.88 ... 74.88 75.12 75.38
  * latitude   (latitude) float64 25.38 25.62 25.88 26.12 ... 74.88 75.12 75.38
Data variables:
    rr         (time, latitude, longitude) float32 nan nan nan ... nan nan nan
```

Here I define the attributes, that xarray uses when plotting

[6]:
```
EOBS['rr'].attrs = {'long_name': 'rainfall',  ##Define the name
 'units': 'mm/day', ## unit
 'standard_name': 'thickness_of_rainfall_amount'} ## original name, not used
EOBS['rr'].mean('time').plot() ## and show the 1950-2019 average February␣
↪precipitation
```

```
/soge-home/users/cenv0732/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/xarray/
↪core/nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

[6]:
```
<matplotlib.collections.QuadMesh at 0x7f9dd1819f10>
```



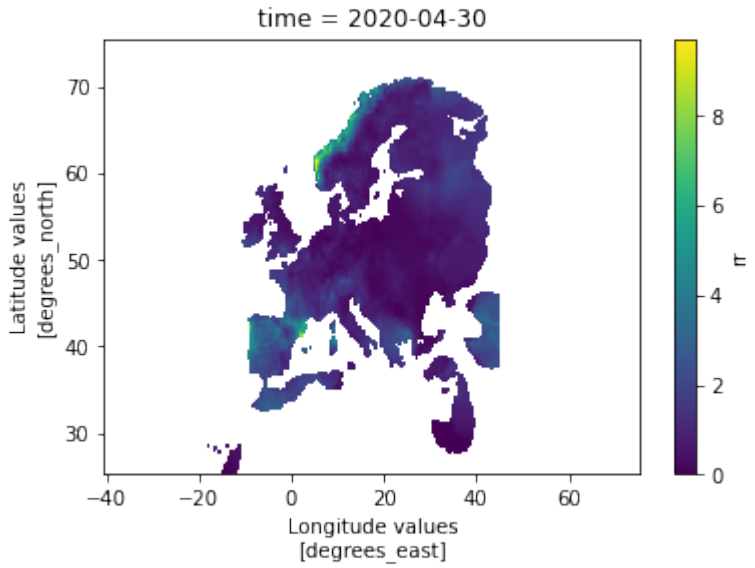
The 2020 data file is separate and needs the same preprocessing:

[8]:
```
EOBS2020 = xr.open_dataset('../UK_example/EOBS/rr_0.25deg_day_2020_grid_ensmean.nc.1
↪') #open
EOBS2020 = EOBS2020.resample(time='1m').mean() #Monthly mean
EOBS2020['rr'].sel(time='2020-04').plot() #show map
EOBS2020 ## display dataset
```

```
/soge-home/users/cenv0732/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/xarray/
↪core/nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

[8]:
```
<matplotlib.collections.QuadMesh at 0x7f9dd15e5820>
```

[8]:
```
<xarray.Dataset>
Dimensions:    (latitude: 201, longitude: 464, time: 12)
Coordinates:
  * time       (time) datetime64[ns] 2020-01-31 2020-02-29 ... 2020-12-31
  * longitude  (longitude) float64 -40.38 -40.12 -39.88 ... 74.88 75.12 75.38
  * latitude   (latitude) float64 25.38 25.62 25.88 26.12 ... 74.88 75.12 75.38
Data variables:
    rr         (time, latitude, longitude) float32 nan nan nan ... nan nan nan
```



We had to download EOBS in two separate files to also obtain the 2020 data. Below, we concatenate the two files and store the processed data for easy import in the future

[10]:
```
EOBS_concat = xr.concat([EOBS,EOBS2020.sel(time='2020-02')],dim='time') ##␣
↪Concatenate the 1950-2019 and 2020 datasets.
EOBS_concat.to_netcdf('../UK_example/EOBS/EOBS.nc') ## And store the 1950-2010␣
↪February precipitation into one nc for future import
```

We then extract UK averaged precipitation SEAS5 and EOBS. We **upscale EOBS** to the SEAS5 grid and apply the same **UK mask** to extract the UK average for both datasets. See Using EOBS + upscaling for an example how to regrid and how to extract a country average timeseries.

Evaluate

Note

From here onward we use R and not python!

We switch to R since we believe R has a better functionality in extreme value statistics.

---

[1]:
```
setwd('../../..')
# getwd()
EOBS_UK_weighted_df <- read.csv("Data/EOBS_UK_weighted_upscaled.csv",␣
↪stringsAsFactors=FALSE)
```

(continues on next page)

(continued from previous page)

```
SEAS5_UK_weighted_df <- read.csv("Data/SEAS5_UK_weighted_masked.csv",␣
↪stringsAsFactors=FALSE)

## Convert the time class to Date format
EOBS_UK_weighted_df$time <- lubridate::ymd(EOBS_UK_weighted_df$time)
str(EOBS_UK_weighted_df)

EOBS_UK_weighted_df_hindcast <- EOBS_UK_weighted_df[
    EOBS_UK_weighted_df$time > '1982-02-01' &
    EOBS_UK_weighted_df$time < '2017-02-01',
    ]


SEAS5_UK_weighted_df$time <- lubridate::ymd(SEAS5_UK_weighted_df$time)
str(SEAS5_UK_weighted_df)
```

```
'data.frame':   71 obs. of  2 variables:
 $ time: Date, format: "1950-02-28" "1951-02-28" ...
 $ rr  : num  4.13 3.25 1.07 1.59 2.59 ...
'data.frame':   4375 obs. of  4 variables:
 $ leadtime: int  2 2 2 2 2 2 2 2 2 2 ...
 $ number  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ time    : Date, format: "1982-02-01" "1983-02-01" ...
 $ tprate  : num  1.62 2.93 3.27 2 3.31 ...
```

Is the UNSEEN ensemble realistic?

To answer this question, we perform three statistical tests: independence, model stability and model fidelity tests. These statistical tests are available through the UNSEEN R package. See evaluation for more info.

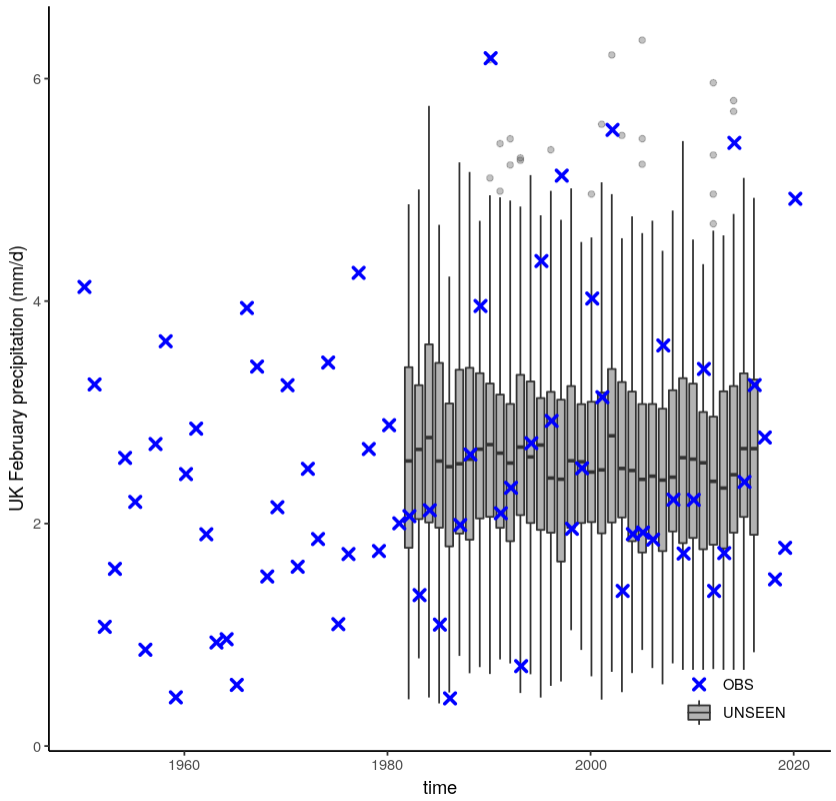
[2]:
```
require(UNSEEN)
```

```
Loading required package: UNSEEN
```

Timeseries

We plot the timeseries of SEAS5 (UNSEEN) and EOBS (OBS) for UK February precipitation.

[3]:
```
unseen_timeseries(ensemble = SEAS5_UK_weighted_df,
                  obs = EOBS_UK_weighted_df,
                  ylab = 'UK February precipitation (mm/d)')
```



We select the timeseries for the **hindcast years 1981-2016**.

[4]:
```
timeseries <- unseen_timeseries(ensemble = SEAS5_UK_weighted_df,
                  obs = EOBS_UK_weighted_df_hindcast,
                  ylab = 'UK February precipitation (mm/d)')
ggsave(timeseries, height = 5, width = 6,    filename = "graphs/UK_timeseries.png")
```

```
Error in ggsave(timeseries, height = 5, width = 6, filename = "graphs/UK_timeseries.
↪png"): could not find function "ggsave"
Traceback:
```

Evaluation tests

With the hindcast dataset we evaluate the independence, stability and fidelity.

First the *independence test*. This test checks if the forecasts are independent. If they are not, the event are not unique and care should be taken in the extreme value analysis. Because of the chaotic behaviour of the atmosphere, independence of precipitation events is expected beyond a lead time of two weeks. Here we use lead times 2-6 months and find that the boxplots are within the expected range (perhaps very small dependence in lead time 2). More info in our paper: https://doi.org/10.31223/osf.io/hyxeq.

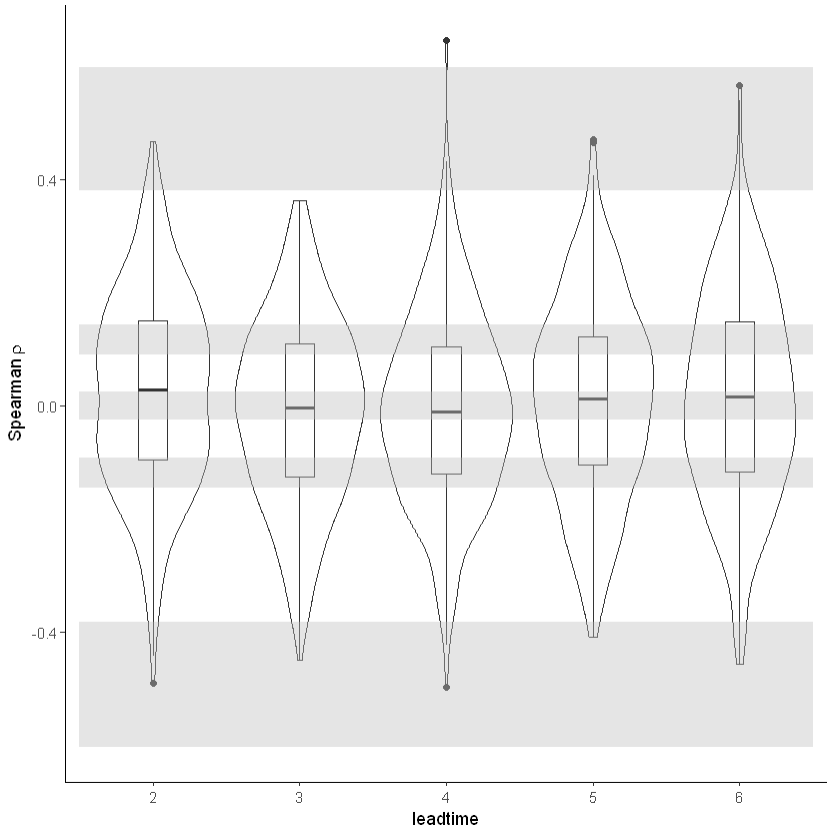
[5]:
```
independence_test(ensemble = SEAS5_UK)
```

```
Warning message:
"Removed 1625 rows containing non-finite values (stat_ydensity)."
```

(continues on next page)

(continued from previous page)

```
Warning message:
"Removed 1625 rows containing non-finite values (stat_boxplot)."
```

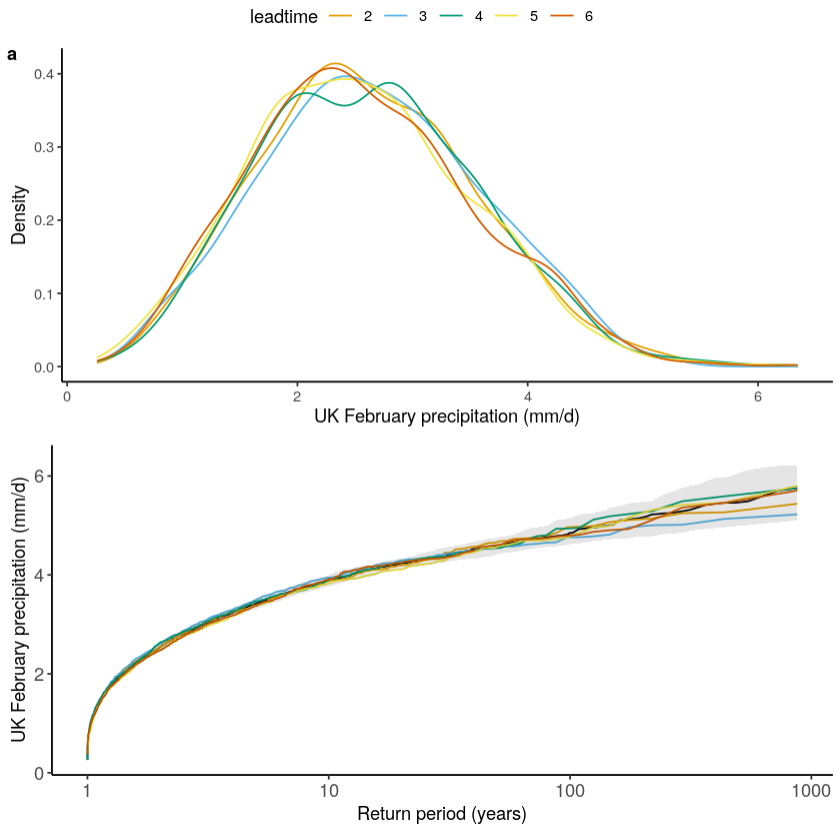


The test for *model stability*: Is there a drift in the simulated precipitation over lead times?

We find that the model is stable for UK February precipitation.

[8]:
```
stability_test(ensemble = SEAS5_UK, lab = 'UK February precipitation (mm/d)')
```

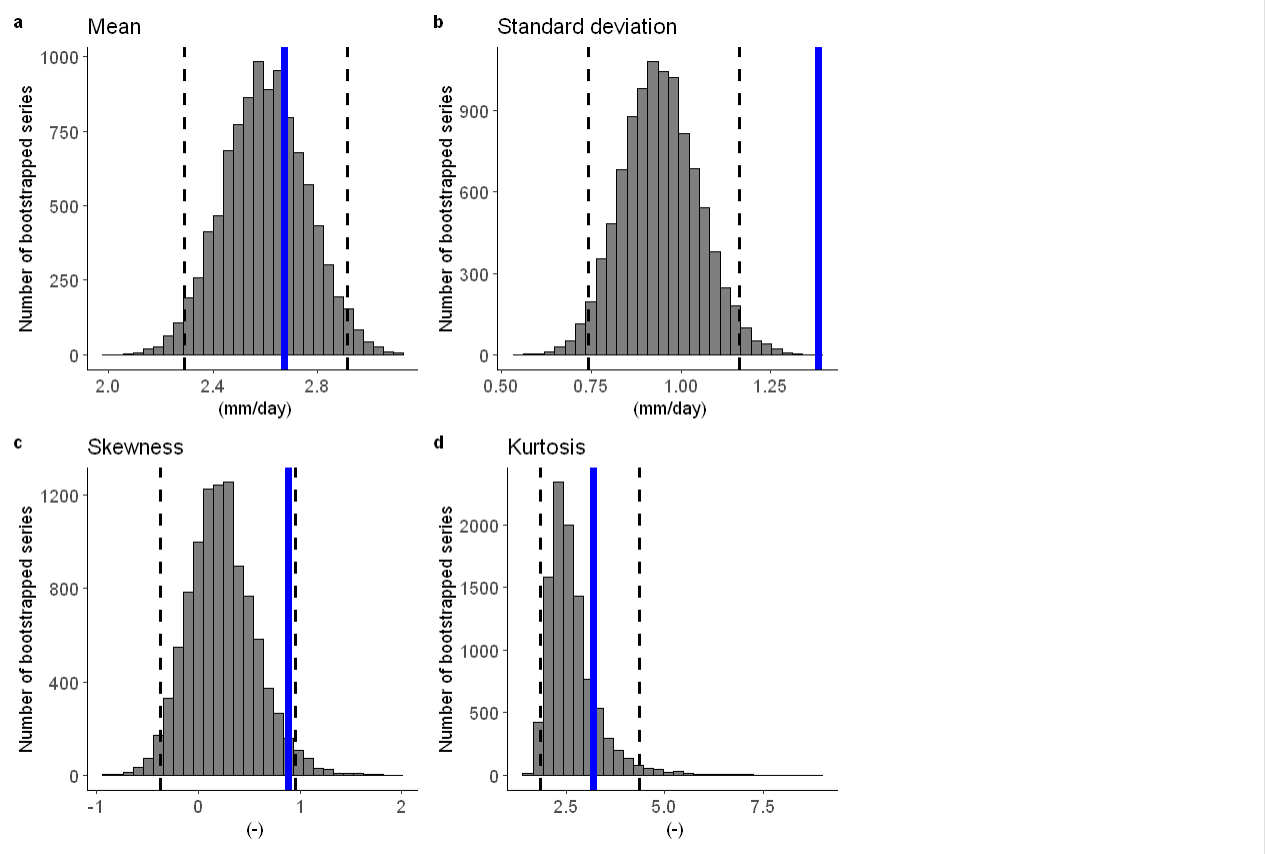
```
Warning message:
“Removed 4 row(s) containing missing values (geom_path).”
```



The *fidelity test* shows us how consistent the model simulations of UNSEEN (SEAS5) are with the observed (EOBS). With this test we can asses systematic biases. The UNSEEN dataset is much larger than the observed – hence they cannot simply be compared. For example, what if we had faced a few more or a few less precipitation extremes purely by chance?

This would influence the observed mean, but not so much influence the UNSEEN ensemble because of the large data sample. Therefore we express the UNSEEN ensemble as a range of plausible means, for data samples of the same length as the observed. We do the same for higher order statistical moments.

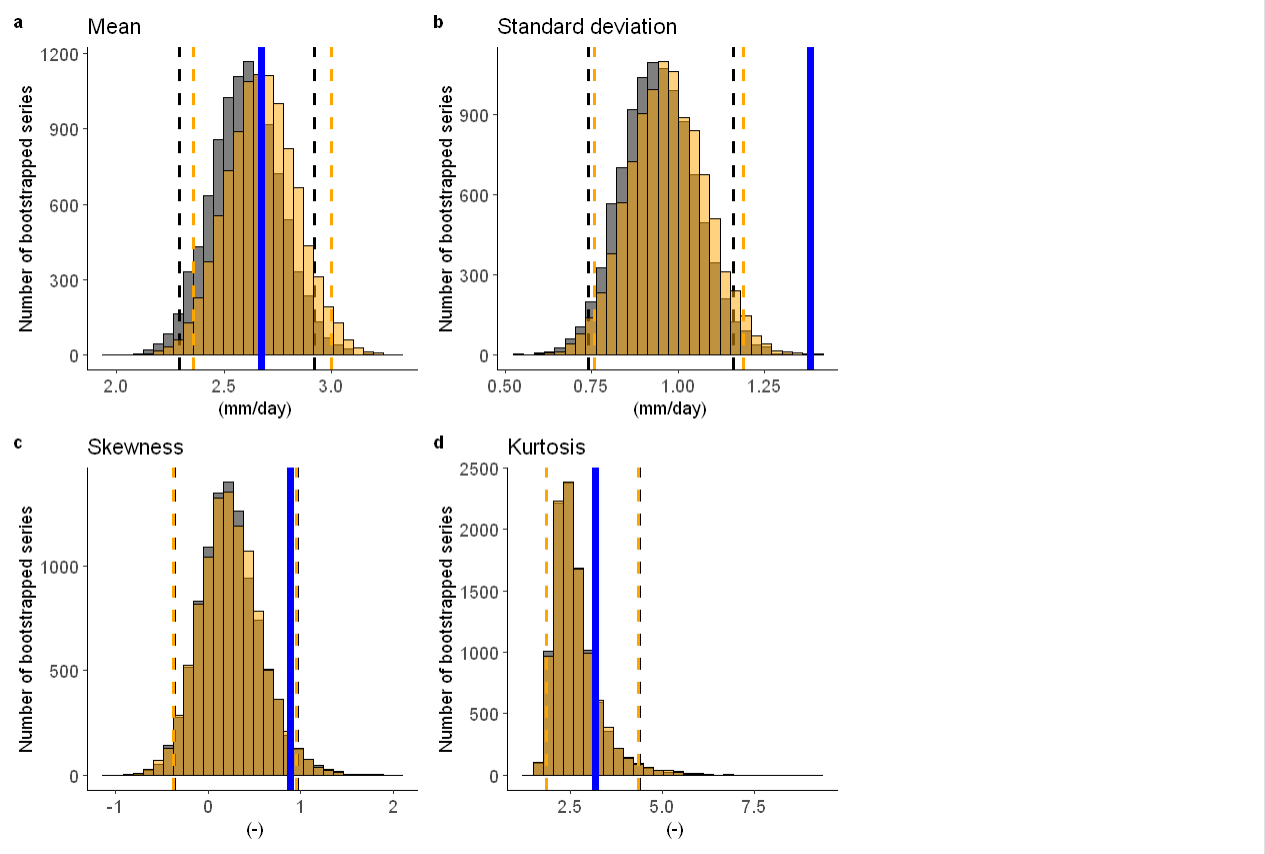
[6]:
```
fidelity_test(obs = EOBS_UK_weighted_df_hindcast$rr,
              ensemble = SEAS5_UK_weighted_df$tprate
             )
```



We find that the standard deviation within the model (the grey histograms and lines) are too low compared to the observed.

We can include a simple mean-bias correction (ratio) in this plot by setting `biascor = TRUE`. However, in this case it won't help:

[7]:
```
fidelity_test(obs = EOBS_UK_weighted_df_hindcast$rr,
              ensemble = SEAS5_UK_weighted_df$tprate,
              biascor = TRUE
             )
```



Check the documentation of the test `?fidelity_test`

Illustrate

[8]:
```
source('src/evt_plot.r')
```

```
Loading required package: Lmoments

Loading required package: distillery


Attaching package: 'extRemes'


The following objects are masked from 'package:stats':

    qqnorm, qqplot
```

First, we fit a Gumbel and a GEV distribution (including shape parameter) to the observed extremes. The Gumbel distribution best describes the data because the p-value of 0.9 is much above 0.05 (based on the likelihood ratio test).

[9]:
```
fit_obs_Gumbel <- fevd(x = EOBS_UK_weighted_df_hindcast$rr,
                       type = "Gumbel"
                      )
```

(continues on next page)

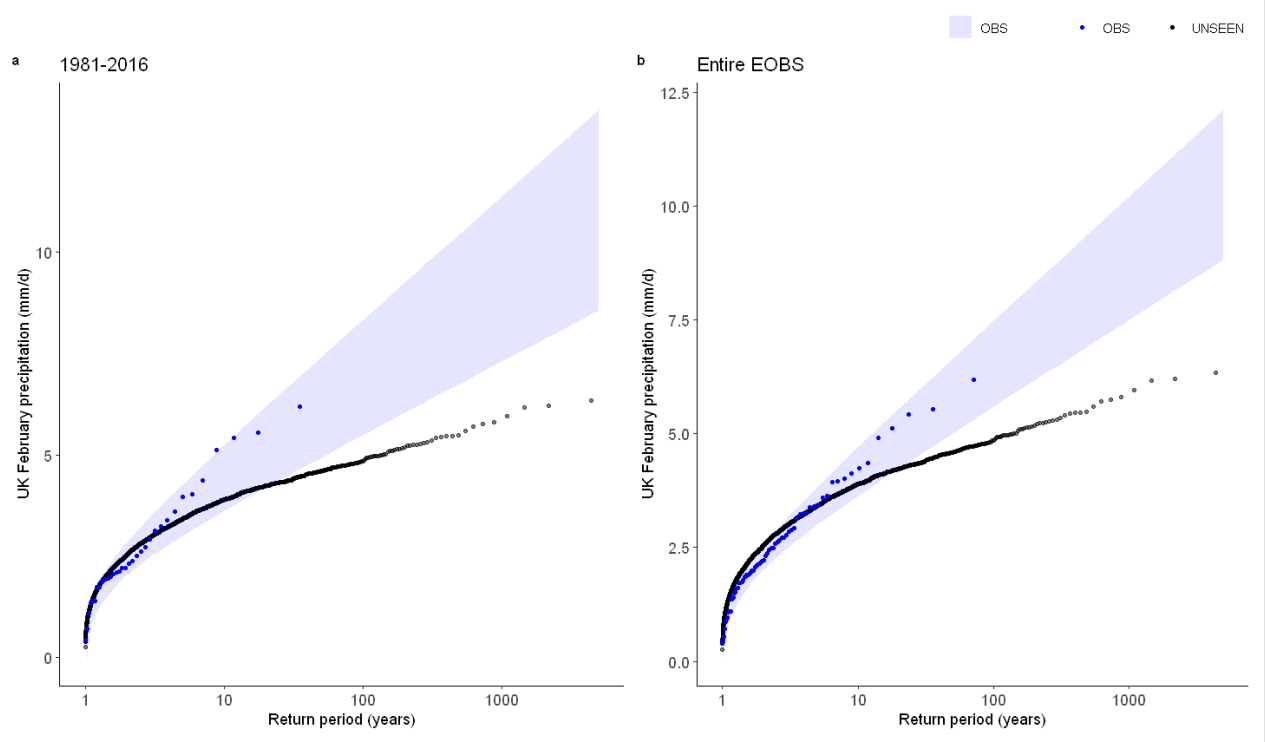
(continued from previous page)

```
fit_obs_GEV <- fevd(x = EOBS_UK_weighted_df_hindcast$rr,
                    type = "GEV"
                   )
lr.test(fit_obs_Gumbel, fit_obs_GEV)
```

```
        Likelihood-ratio Test

data:  EOBS_UK_weighted_df_hindcast$rrEOBS_UK_weighted_df_hindcast$rr
Likelihood-ratio = 0.014629, chi-square critical value = 3.8415, alpha
= 0.0500, Degrees of Freedom = 1.0000, p-value = 0.9037
alternative hypothesis: greater
```

We show the gumbel plot for the observed (EOBS) and UNSEEN (SEAS5 hindcast data). This shows that the UNSEEN simulations are not within the uncertainty range of the observations. This has to do with the variability of the model that is too low, as indicated in the evaluation section.

[13]:
```
options(repr.plot.width = 12)
Gumbel_hindcast <- EVT_plot(ensemble = SEAS5_UK_weighted_df$tprate,
                            obs = EOBS_UK_weighted_df_hindcast$rr,
                            main = "1981-2016",
                            GEV_type = "Gumbel",
#                             ylim = 3,
                            y_lab = 'UK February precipitation (mm/d)'
                           )
GEV_hindcast <- EVT_plot(ensemble = SEAS5_UK_weighted_df$tprate,
                                    obs = EOBS_UK_weighted_df$rr,
                                    main = "Entire EOBS",
                                    GEV_type = "Gumbel",
#                                     ylim = 3,
                                    y_lab = 'UK February precipitation (mm/d)'
                                   )

ggarrange(Gumbel_hindcast, GEV_hindcast,
  labels = c("a", "b"), # ,"c","d"),
  common.legend = T,
  font.label = list(size = 10, color = "black", face = "bold", family = NULL),
  ncol = 2, nrow = 1
)
```



Why is there too little variability within UK february simulations?

This can be fed back to model developers to help improve the models.

We could further explore the use of other observational datasets and other model simulations.

1.4 Retrieve

We want to download the monthly precipitation for February. I use the automatically generated request from the CDS server. There are two datasets we can use to download the data: Seasonal forecast daily data on single levels and Seasonal forecast monthly statistics on single levels. We will use the latter for easy downloading of the monthly values. If we want to go to higher temporal resolution, such as daily extremes, we will have to consult the other dataset.

To get started with CDS, you have to register at https://cds.climate.copernicus.eu/ and copy your UID and API key from https://cds.climate.copernicus.eu/user in the `~/.cdsapirc` file in the home directory of your user. See the ml-flood project for more details

[7]:
```
UID = 'UID'
API_key = 'API_key'
```

[8]:
```
import os
#Uncomment the following lines to write the UID and API key in the .cdsapirc file
# with open(os.path.join(os.path.expanduser('~'), '.cdsapirc'), 'w') as f:
#     f.write('url: https://cds.climate.copernicus.eu/api/v2\n')
#     f.write(f'key: {UID}:{API_key}')
```

[8]:
```
46
```

[8]:
```
47
```

1.4.1 Import packages

[1]:
```
##This is so variables get printed within jupyter
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

[2]:
```
##import packages
import os
import cdsapi ## check the current working directory, which should be the UNSEEN-open␣
↪directory
import numpy as np
import xarray as xr
import matplotlib.pyplot as plt
import numpy as np
import cartopy
import cartopy.crs as ccrs
```

[3]:
```
##We want the working directory to be the UNSEEN-open directory
pwd = os.getcwd() ##current working directory is UNSEEN-open/Notebooks/1.Download
pwd #print the present working directory
os.chdir(pwd+'/../../') # Change the working directory to UNSEEN-open
os.getcwd() #print the working directory
```

[3]:
```
'C:\\Users\\Timo\\OneDrive - Loughborough University\\GitHub\\UNSEEN-open\\doc\\
↪Notebooks\\1.Download'
```

[3]:
```
'C:\\Users\\Timo\\OneDrive - Loughborough University\\GitHub\\UNSEEN-open\\doc'
```

1.4.2 First download

In our request, we will use the monthly mean. Interestingly, there is also the option to use the monthly maximum! We previously downloaded the data on daily resolution and extracted the monthly (or seasonal) maximum from that data. If we could just download the monthly maximum instead that might save a lot of processing power! However, you would be restricted to daily extremes only, for multi-day extremes (5 days is often used), you would have to do the original processing workflow. We select the UK domain to reduce the size of the download.

Here I download the monthly mean total precipitation (both convective and large scale precipitation) forecast for February 1993. It downloads all 25 ensemble members for the forecasts initialized in january.

[4]:
```
##Our first download:

c = cdsapi.Client()

c.retrieve(
    'seasonal-monthly-single-levels',
    {
        'format': 'netcdf',
        'originating_centre': 'ecmwf',
        'system': '5',
        'variable': 'total_precipitation',
        'product_type': [
            'monthly_mean', #'monthly_maximum',, 'monthly_standard_deviation',
        ],
        'year': '1993', #data before 1993 is available.
        'month': '01', #Initialization month. Target month is February (2),␣
↪initialization months are August-January (8-12,1)
```

(continues on next page)

(continued from previous page)

```
        'leadtime_month': [ ##Use of single months is much faster. Leadtime 0 does␣
↪not exist. The first lead time is 1.
            '1', '2',
        ],
        'area': [##Select UK domain to reduce the size of the download
            60, -11, 50,
            2,
        ],
    },
    'Data/First_download.nc') ##can I use nc? yes!
```

```
2020-05-13 10:08:56,140 INFO Welcome to the CDS
2020-05-13 10:08:56,142 INFO Sending request to https://cds.climate.copernicus.eu/api/
↪v2/resources/seasonal-monthly-single-levels
2020-05-13 10:08:56,983 INFO Request is completed
2020-05-13 10:08:56,984 INFO Downloading http://136.156.132.110/cache-compute-0001/
↪cache/data0/adaptor.mars.external-1589266964.5635436-26283-29-a38e8975-b0ec-49ee-
↪8f9b-7dea389f59cf.nc to Data/First_download.nc (16.4K)
2020-05-13 10:08:57,131 INFO Download rate 112.7K/s
```

[4]:
```
Result(content_length=16800,content_type=application/x-netcdf,location=http://136.156.
↪132.110/cache-compute-0001/cache/data0/adaptor.mars.external-1589266964.5635436-
↪26283-29-a38e8975-b0ec-49ee-8f9b-7dea389f59cf.nc)
```

Use xarray to visualize the netcdf file

I open the downloaded file and plot February 1993 precipitation over the UK.

[5]:
```
pr_1993_ds=xr.open_dataset('Data/First_download.nc')
pr_1993_ds
```

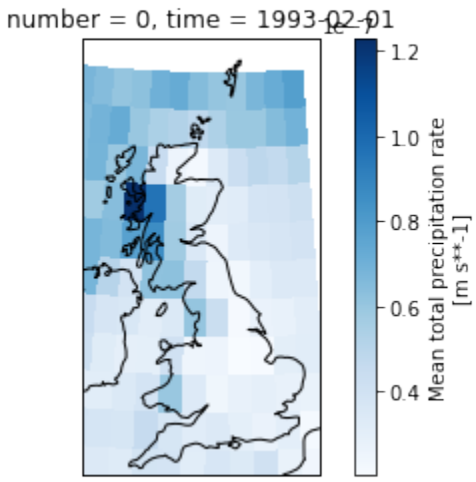
[5]:
```
<xarray.Dataset>
Dimensions:    (latitude: 11, longitude: 14, number: 25, time: 2)
Coordinates:
  * longitude  (longitude) float32 -11.0 -10.0 -9.0 -8.0 ... -1.0 0.0 1.0 2.0
  * latitude   (latitude) float32 60.0 59.0 58.0 57.0 ... 53.0 52.0 51.0 50.0
  * number     (number) int32 0 1 2 3 4 5 6 7 8 9 ... 16 17 18 19 20 21 22 23 24
  * time       (time) datetime64[ns] 1993-01-01 1993-02-01
Data variables:
    tprate     (time, number, latitude, longitude) float32 ...
Attributes:
    Conventions:  CF-1.6
    history:      2020-05-12 07:02:45 GMT by grib_to_netcdf-2.16.0: /opt/ecmw...
```

I select ensemble member 0 and february precipitation ('tprate' called apparently) and I use cartopy to make the map.

[6]:
```
## Use cartopy for nicer maps
ax = plt.axes(projection= ccrs.OSGB())
pr_1993_ds['tprate'].sel(number=0,time='1993-02').plot(transform=ccrs.PlateCarree(),
↪cmap=plt.cm.Blues, ax=ax)  #,cmap=plt.cm.Blues,

# ax.set_extent(extent)
ax.coastlines(resolution='50m')
plt.draw()
```

[6]:
```
<matplotlib.collections.QuadMesh at 0x7f5435adfa60>
```

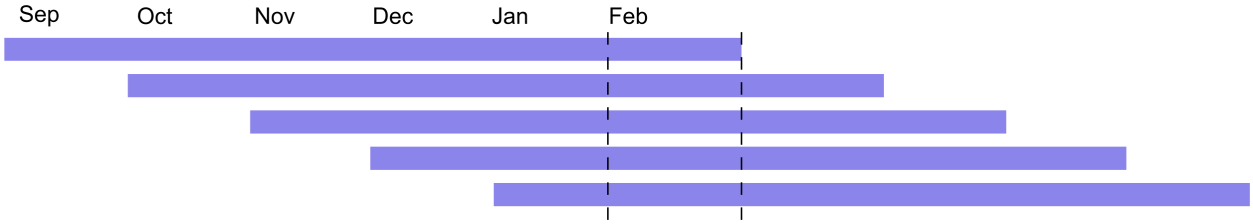
[6]:
```
<cartopy.mpl.feature_artist.FeatureArtist at 0x7f5435b6ff10>
```



Download all data

We will be using the SEAS5 hindcast, which is a dataset running from 1981-2016. The hindcast is initialized every month with 25 ensemble members and the forecast run for 6 months, indicated by blue horizontal bars below. February is forecasted by 6 initialization months (September-February). We discard the first month of the forecast because of dependence between the forecasts, explained in the evaluation section and are left with 5 initialization months (Sep-Jan) and 25 ensemble members forecasting february precipitation each year, totalling to an increase of 125 times the observed length.

For a summary of all available C3S seasonal hindcasts, their initialization months and more specifics, please see ECMWF page and the SEAS5 paper.



The first download example above downloaded all 25 ensemble members for the forecast initialized in January (the bottom bar). We should repeat this over the other initialization month and over all years (1981-2016).

[58]:
```
init_months = np.append(np.arange(9,13),1) ## Initialization months 9-12,1 (Sep-Jan)
init_months
years = np.arange(1982,2017)
years
```

[58]:
```
array([ 9, 10, 11, 12,  1])
```

[58]:
```
array([1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992,
       1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003,
       2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014,
       2015, 2016])
```

For our download, we loop over initialization months and years. Because we only want February precipitation, the leadtime month (forecast length) changes with the initialization month. For example, in the September initialized forecasts, we only want the leadtime month 6 = February. For August initialized this is leadtime 5, etc. Furthermore, the year the forecast is initialized is required for the download. For September - December initialized forecasts, this is the target year-1. For January it is the same year as the target year. For example, for the first two target years this looks like the following:

[101]:
```
for j in range(2):#len(years)):
    for i in range(len(init_months)):
        init_month = init_months[i]
        leadtime_month = 6-i
        if init_month == 1:
            year = years[j]
        else:
            year = years[j]-1
        print ('year = ' + str(year) +' init_month = ' + str(init_month) + ' leadtime_
↪month = ' + str(leadtime_month))
```

```
year = 1981 init_month = 9 leadtime_month = 6
year = 1981 init_month = 10 leadtime_month = 5
year = 1981 init_month = 11 leadtime_month = 4
year = 1981 init_month = 12 leadtime_month = 3
year = 1982 init_month = 1 leadtime_month = 2
year = 1982 init_month = 9 leadtime_month = 6
year = 1982 init_month = 10 leadtime_month = 5
year = 1982 init_month = 11 leadtime_month = 4
year = 1982 init_month = 12 leadtime_month = 3
year = 1983 init_month = 1 leadtime_month = 2
```

Write a function that is used for the download.

[72]:
```
def retrieve(variable, originating_centre, year, init_month, leadtime_month):

    c.retrieve(
        'seasonal-monthly-single-levels',
        {
            'format': 'netcdf',
            'originating_centre': originating_centre,
            'system': '5',
            'variable': variable,
            'product_type': [
                'monthly_mean', #'monthly_maximum',, 'monthly_standard_deviation',
            ],
            'year': str(year), #data before 1993 is available.
            'month': "%.2i" % init_month, #Initialization month. Target month is␣
↪February (2), initialization months are August-January (8-12,1)
            'leadtime_month': [ ##The lead times you want. Use of single months is␣
↪much faster. Leadtime 0 does not exist. The first lead time is 1.
                #For initialization month 1 (January), the leadtime months is 2␣
↪(February). For initialization month 12 (december), the lead time month is 3␣
↪(February).
                str(leadtime_month),
            ],
            'area': [##Select UK domain to reduce the size of the download
                     ## 25N-75N x. 40W-75E
                60, -11, 50, 2,
            ],
```

(continues on next page)

(continued from previous page)

```
        },
        '../UK_example/'+ str(year) + "%.2i" % init_month + '.nc')

# retrieve(variable = 'total_precipitation',originating_centre = 'ecmwf', year =␣
↪years[0], init_month = "%.2i" % init_months[0])
```

And start the download! In total, we request 35 years x initialization dates = 175 requests. I could try sending just 5 request of the different initialization dates for all years?

[ ]:
```
for j in range(len(years)):  ##add if error still continue
    for i in range(len(init_months)):
        init_month = init_months[i]
        leadtime_month = 6 - i
        if init_month == 1:
            year = years[j]
        else:
            year = years[j] - 1
        retrieve(variable='total_precipitation',
                 originating_centre='ecmwf',
                 year=year,
                 init_month=init_month,
                 leadtime_month=leadtime_month)
```

```
2020-05-18 10:14:48,767 INFO Welcome to the CDS
2020-05-18 10:14:48,768 INFO Sending request to https://cds.climate.copernicus.eu/api/
↪v2/resources/seasonal-monthly-single-levels
2020-05-18 10:14:49,485 INFO Downloading http://136.156.132.235/cache-compute-0006/
↪cache/data5/adaptor.mars.external-1589380912.7108843-4209-7-1add31ae-a0cd-44ce-83ac-
↪9ff7c97f1b01.nc to ../UK_example/198109.nc (8.9K)
2020-05-18 10:14:49,575 INFO Download rate 101.5K/s
2020-05-18 10:14:49,803 INFO Welcome to the CDS
2020-05-18 10:14:49,804 INFO Sending request to https://cds.climate.copernicus.eu/api/
↪v2/resources/seasonal-monthly-single-levels
2020-05-18 10:14:50,498 INFO Downloading http://136.156.132.153/cache-compute-0002/
↪cache/data4/adaptor.mars.external-1589381056.172494-12462-1-c9714216-87ac-49bc-be19-
↪260627a9077d.nc to ../UK_example/198110.nc (8.9K)
2020-05-18 10:14:50,571 INFO Download rate 124.6K/s
2020-05-18 10:14:51,070 INFO Welcome to the CDS
2020-05-18 10:14:51,071 INFO Sending request to https://cds.climate.copernicus.eu/api/
↪v2/resources/seasonal-monthly-single-levels
2020-05-18 10:14:51,213 INFO Downloading http://136.156.132.235/cache-compute-0006/
↪cache/data9/adaptor.mars.external-1589381301.6300867-8112-3-49ba0ab2-34fe-4364-9dec-
↪700bf911b079.nc to ../UK_example/198111.nc (8.9K)
2020-05-18 10:14:51,254 INFO Download rate 219.7K/s
2020-05-18 10:14:51,415 INFO Welcome to the CDS
2020-05-18 10:14:51,416 INFO Sending request to https://cds.climate.copernicus.eu/api/
↪v2/resources/seasonal-monthly-single-levels
2020-05-18 10:14:51,548 INFO Request is queued
```

The download sometimes fails. When redoing the request it does download. I don't know what is causing the failure? Below I donwload the file that failed.

[97]:
```
#201501 missing

year = 2015
init_month = 1
leadtime_month = 2
```

(continues on next page)

(continued from previous page)

```
retrieve(variable = 'total_precipitation',originating_centre = 'ecmwf', year = year,
                 init_month = init_month, leadtime_month = leadtime_month)
```

```
2020-05-15 11:51:16,127 INFO Welcome to the CDS
2020-05-15 11:51:16,129 INFO Sending request to https://cds.climate.copernicus.eu/api/
↪v2/resources/seasonal-monthly-single-levels
2020-05-15 11:51:16,327 INFO Downloading http://136.156.133.46/cache-compute-0015/
↪cache/data7/adaptor.mars.external-1589527607.2123153-8094-37-3b786f72-2e2a-462f-
↪bbb8-9c8d89c05102.nc to ../UK_example/201501.nc (8.9K)
2020-05-15 11:51:16,485 INFO Download rate 56.7K/s
```

Retrieve function

We have written a module where the above procedure is done automatically. Here we load the retrieve module and retrieve SEAS5 and ERA5 data for the examples by selecting the variable, target month(s), area and folder where we want to download the file in.

The main function to download the data is `retrieve.retrieve_SEAS5`. The function only downloads the target months, for each year and each intialization month. To do this, it obtains the initialization months and leadtimes from the selected target month(s). For the UK example, we select February as our target month, hence sep-jan will be our initialization months with leadtimes 2-6, see *Download all*.

[7]:
```
retrieve.print_arguments([2])
```

```
year = 1982 init_month = 1 leadtime_month = [2]
year = 1981 init_month = 12 leadtime_month = [3]
year = 1981 init_month = 11 leadtime_month = [4]
year = 1981 init_month = 10 leadtime_month = [5]
year = 1981 init_month = 9 leadtime_month = [6]
```

For the Siberia example this will be different, since the target months are march-may:

[8]:
```
retrieve.print_arguments([3,4,5])
```

```
year = 1982 init_month = 2 leadtime_month = [2 3 4]
year = 1982 init_month = 1 leadtime_month = [3 4 5]
year = 1981 init_month = 12 leadtime_month = [4 5 6]
```

Call `?retrieve.retrieve_SEAS5` to see the documentation.

For the California example, we use:

[ ]:
```
retrieve.retrieve_SEAS5(
    variables=['2m_temperature', '2m_dewpoint_temperature'],
    target_months=[8],
    area=[70, -130, 20, -70],
    years=np.arange(1981, 2021),
    folder='E:/PhD/California_example/SEAS5/')
```

[ ]:
```
retrieve.retrieve_ERA5(variables=['2m_temperature', '2m_dewpoint_temperature'],
                       target_months=[8],
                       area=[70, -130, 20, -70],
                       folder='E:/PhD/California_example/SEAS5/')
```

For the Siberia example:

```
[ ]: retrieve.retrieve_SEAS5(
         variables=['2m_temperature', '2m_dewpoint_temperature'],
         target_months=[3, 4, 5],
         area=[70, -11, 30, 120],
         years=np.arange(1981, 2021),
         folder='../Siberia_example/SEAS5/')
```

```
[ ]: retrieve.retrieve_ERA5(variables = ['2m_temperature','2m_dewpoint_temperature'],
                           target_months = [3,4,5],
                           area = [70, -11, 30, 120],
                           folder = '../Siberia_example/ERA5/')
```

And for the UK example:

```
[ ]: retrieve.retrieve_SEAS5(variables = 'total_precipitation',
                            target_months = [2],
                            area = [60, -11, 50, 2],
                            folder = '../UK_example/SEAS5/')
```

```
[ ]: retrieve.retrieve_ERA5(variables = 'total_precipitation',
                           target_months = [2],
                           area = [60, -11, 50, 2],
                           folder = '../UK_example/ERA5/')
```

EOBS data download

I tried to download EOBS through CDS, but the Product is temporally disabled for maintenance purposes (see below). As workaround I downloaded EOBS (from 1950 - 2019) and the most recent EOBS data (2020) here. Note, you have to register as E-OBS user.

```
[99]: c.retrieve(
          'insitu-gridded-observations-europe',
          {
              'version': 'v20.0e',
              'format': 'zip',
              'product_type': 'ensemble_mean',
              'variable': 'precipitation_amount',
              'grid_resolution': '0_25',
              'period': 'full_period',
          },
          '../UK_example/EOBS/EOBS.zip')
```

```
2020-05-15 14:06:44,721 INFO Welcome to the CDS
2020-05-15 14:06:44,722 INFO Sending request to https://cds.climate.copernicus.eu/api/
↪v2/resources/insitu-gridded-observations-europe
```

```
---------------------------------------------------------------------------
HTTPError                                 Traceback (most recent call last)
~/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/cdsapi/api.py in _api(self, url,
↪ request, method)
    388         try:
--> 389             result.raise_for_status()
    390             reply = result.json()

~/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/requests/models.py in raise_for_
↪status(self)
```

(continues on next page)

(continued from previous page)

```
    940         if http_error_msg:
--> 941             raise HTTPError(http_error_msg, response=self)
    942

HTTPError: 403 Client Error:  for url: https://cds.climate.copernicus.eu/api/v2/
↪resources/insitu-gridded-observations-europe

During handling of the above exception, another exception occurred:

Exception                                 Traceback (most recent call last)
<ipython-input-99-d12768b41b79> in <module>
----> 1 c.retrieve(
      2     'insitu-gridded-observations-europe',
      3     {
      4         'version': 'v20.0e',
      5         'format': 'zip',

~/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/cdsapi/api.py in retrieve(self,␣
↪name, request, target)
    315
    316     def retrieve(self, name, request, target=None):
--> 317         result = self._api('%s/resources/%s' % (self.url, name), request,
↪'POST')
    318         if target is not None:
    319             result.download(target)

~/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/cdsapi/api.py in _api(self, url,
↪ request, method)
    408                                     "of '%s' at %s" % (t['title'], t['url']))
    409                     error = '. '.join(e)
--> 410                 raise Exception(error)
    411             else:
    412                 raise

Exception: Product temporally disabled for maintenance purposes. Sorry for the␣
↪inconvenience, please try again later.
```

1.5 Preprocess

The preprocessing steps consist of merging all retrieved files into one xarray dataset and extracting the spatial and temporal average of the event of interest.

1.5.1 Merge

Here it is shown how all retrieved files are loaded into one xarray dataset, for both SEAS5 and for ERA5.

SEAS5

All retrieved seasonal forecasts are loaded into one xarray dataset. The amount of files retrieved depends on the temporal extent of the extreme event that is being analyzed (i.e are you looking at a monthly average or a seasonal average?). For the Siberian heatwave, we have retrieved 105 files (one for each of the 35 years and for each of the three lead times, (*see Retrieve*). For the UK, we are able to use more forecasts, because the target month is shorter: one month as compared to three months for the Siberian example. We retrieved 5 leadtimes x 35 = 175 files.

Each netcdf file contains 25 ensemble members, hence has the dimensions lat, lon, number (25 ensembles). Here we create an xarray dataset that also contains the dimensions time (35 years) and leadtime (5 initialization months). To generate this, we loop over lead times, and open all 35 years of the lead time and then concatenate those leadtimes.

[1]:
```
##This is so variables get printed within jupyter
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

[2]:
```
import os
import sys
sys.path.insert(0, os.path.abspath('../../../'))
import src.cdsretrieve as retrieve
```

[3]:
```
os.chdir(os.path.abspath('../../../'))
os.getcwd() #print the working directory
```

[3]:
```
'C:\\Users\\Timo\\OneDrive - Loughborough University\\GitHub\\UNSEEN-open'
```

[4]:
```
import xarray as xr
import numpy as np

def merge_SEAS5(folder, target_months):
    init_months, leadtimes = retrieve._get_init_months(target_months)
    print('Lead time: ' + "%.2i" % init_months[0])
    SEAS5_ld1 = xr.open_mfdataset(
        folder + '*' + "%.2i" % init_months[0] + '.nc',
        combine='by_coords')  # Load the first lead time
    SEAS5 = SEAS5_ld1  # Create the xarray dataset to concatenate over
    for init_month in init_months[1:len(init_months)]:  ## Remove the first that we␣
↪already have
        print(init_month)
        SEAS5_ld = xr.open_mfdataset(
            folder + '*' + "%.2i" % init_month + '.nc',
            combine='by_coords')
        SEAS5 = xr.concat([SEAS5, SEAS5_ld], dim='leadtime')
    SEAS5 = SEAS5.assign_coords(leadtime = np.arange(len(init_months)) + 2) # assign␣
↪leadtime coordinates
    return(SEAS5)
```

[ ]:
```
SEAS5_Siberia = merge_SEAS5(folder='../Siberia_example/SEAS5/',
                            target_months=[3, 4, 5])
```

[8]:
```
SEAS5_Siberia
```

[8]:
```
<xarray.Dataset>
Dimensions:    (latitude: 41, leadtime: 3, longitude: 132, number: 51, time: 117)
Coordinates:
  * time       (time) datetime64[ns] 1982-03-01 1982-04-01 ... 2020-05-01
  * latitude   (latitude) float32 70.0 69.0 68.0 67.0 ... 33.0 32.0 31.0 30.0
  * number     (number) int64 0 1 2 3 4 5 6 7 8 9 ... 42 43 44 45 46 47 48 49 50
  * longitude  (longitude) float32 -11.0 -10.0 -9.0 -8.0 ... 118.0 119.0 120.0
  * leadtime   (leadtime) int64 2 3 4
Data variables:
    t2m        (leadtime, time, number, latitude, longitude) float32 dask.array
↪<chunksize=(1, 3, 51, 41, 132), meta=np.ndarray>
    d2m        (leadtime, time, number, latitude, longitude) float32 dask.array
↪<chunksize=(1, 3, 51, 41, 132), meta=np.ndarray>
Attributes:
    Conventions:  CF-1.6
    history:      2020-09-08 09:33:24 GMT by grib_to_netcdf-2.16.0: /opt/ecmw...
```

You can for example select a the lat, long, time, ensemble member and lead time as follows (add `.load()` to see the values):

[ ]:
```
SEAS5_Siberia.sel(latitude=60,
                  longitude=-10,
                  time='2000-03',
                  number=26,
                  leadtime=3).load()
```

We can repeat this for the UK example, where just February is the target month:

[10]:
```
SEAS5_UK = merge_SEAS5(folder = '../UK_example/SEAS5/', target_months = [2])
```

```
Lead time: 01
12
11
10
9
```

The SEAS5 total precipitation rate is in m/s. You can easily convert this and change the attributes. Click on the show/hide attributes button to see the assigned attributes.

[11]:
```
SEAS5_UK['tprate'] = SEAS5_UK['tprate'] * 1000 * 3600 * 24 ## From m/s to mm/d
SEAS5_UK['tprate'].attrs = {'long_name': 'rainfall',
 'units': 'mm/day',
 'standard_name': 'thickness_of_rainfall_amount'}
SEAS5_UK
```

[11]:
```
<xarray.Dataset>
Dimensions:    (latitude: 11, leadtime: 5, longitude: 14, number: 25, time: 35)
Coordinates:
  * time       (time) datetime64[ns] 1982-02-01 1983-02-01 ... 2016-02-01
  * latitude   (latitude) float32 60.0 59.0 58.0 57.0 ... 53.0 52.0 51.0 50.0
  * number     (number) int32 0 1 2 3 4 5 6 7 8 9 ... 16 17 18 19 20 21 22 23 24
  * longitude  (longitude) float32 -11.0 -10.0 -9.0 -8.0 ... -1.0 0.0 1.0 2.0
  * leadtime   (leadtime) int64 2 3 4 5 6
Data variables:
    tprate     (leadtime, time, number, latitude, longitude) float32 dask.array
↪<chunksize=(1, 1, 25, 11, 14), meta=np.ndarray>
Attributes:
```

(continues on next page)

(continued from previous page)

```
    Conventions:  CF-1.6
    history:      2020-05-13 14:49:43 GMT by grib_to_netcdf-2.16.0: /opt/ecmw...
```

ERA5

For each year a netcdf file is downloaded. They are named ERA5_yyyy, for example ERA5_1981. Therefore, we can load ERA5 by combining all downloaded years:

[12]:
```
ERA5_Siberia = xr.open_mfdataset('../Siberia_example/ERA5/ERA5_????.nc',combine='by_
↪coords') ## open the data
ERA5_Siberia
```

[12]:
```
<xarray.Dataset>
Dimensions:    (latitude: 41, longitude: 132, time: 126)
Coordinates:
  * latitude   (latitude) float32 70.0 69.0 68.0 67.0 ... 33.0 32.0 31.0 30.0
  * longitude  (longitude) float32 -11.0 -10.0 -9.0 -8.0 ... 118.0 119.0 120.0
  * time       (time) datetime64[ns] 1979-03-01 1979-04-01 ... 2020-05-01
Data variables:
    t2m        (time, latitude, longitude) float32 dask.array<chunksize=(3, 41, 132),␣
↪meta=np.ndarray>
    d2m        (time, latitude, longitude) float32 dask.array<chunksize=(3, 41, 132),␣
↪meta=np.ndarray>
Attributes:
    Conventions:  CF-1.6
    history:      2020-09-08 13:26:08 GMT by grib_to_netcdf-2.16.0: /opt/ecmw...
```

[13]:
```
ERA5_UK = xr.open_mfdataset('../UK_example/ERA5/ERA5_????.nc',combine='by_coords') ##␣
↪open the data
ERA5_UK
```

[13]:
```
<xarray.Dataset>
Dimensions:    (latitude: 11, longitude: 14, time: 42)
Coordinates:
  * latitude   (latitude) float32 60.0 59.0 58.0 57.0 ... 53.0 52.0 51.0 50.0
  * longitude  (longitude) float32 -11.0 -10.0 -9.0 -8.0 ... -1.0 0.0 1.0 2.0
  * time       (time) datetime64[ns] 1979-02-01 1980-02-01 ... 2020-02-01
Data variables:
    tp         (time, latitude, longitude) float32 dask.array<chunksize=(1, 11, 14),␣
↪meta=np.ndarray>
Attributes:
    Conventions:  CF-1.6
    history:      2020-09-08 13:36:54 GMT by grib_to_netcdf-2.16.0: /opt/ecmw...
```

1.5.2 Event definition

Time selection

For the UK, the event of interest is UK February average precipitation. Since we download monthly averages, we do not have to do any preprocessing along the time dimension here. For the Siberian heatwave, we are interested in the March-May average. Therefore we need to take the seasonal average of the monthly timeseries. We cannot take the simple mean of the three months, because they have a different number of days in the months, see this example. Therefore we take a weighted average:

```
[14]: month_length = SEAS5_Siberia.time.dt.days_in_month
month_length
```

```
[14]: <xarray.DataArray 'days_in_month' (time: 117)>
array([31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30,
       31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31,
       30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31,
       31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30,
       31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31,
       30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31,
       31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31, 31, 30, 31])
Coordinates:
  * time     (time) datetime64[ns] 1982-03-01 1982-04-01 ... 2020-05-01
```

```
[15]: # Calculate the weights by grouping by 'time.season'.
weights = month_length.groupby('time.year') / month_length.groupby('time.year').sum()
weights
```

```
[15]: <xarray.DataArray 'days_in_month' (time: 117)>
array([0.33695652, 0.32608696, 0.33695652, 0.33695652, 0.32608696,
       0.33695652, 0.33695652, 0.32608696, 0.33695652, 0.33695652,
       0.32608696, 0.33695652, 0.33695652, 0.32608696, 0.33695652,
       0.33695652, 0.32608696, 0.33695652, 0.33695652, 0.32608696,
       0.33695652, 0.33695652, 0.32608696, 0.33695652, 0.33695652,
       0.32608696, 0.33695652, 0.33695652, 0.32608696, 0.33695652,
       0.33695652, 0.32608696, 0.33695652, 0.33695652, 0.32608696,
       0.33695652, 0.33695652, 0.32608696, 0.33695652, 0.33695652,
       0.32608696, 0.33695652, 0.33695652, 0.32608696, 0.33695652,
       0.33695652, 0.32608696, 0.33695652, 0.33695652, 0.32608696,
       0.33695652, 0.33695652, 0.32608696, 0.33695652, 0.33695652,
       0.32608696, 0.33695652, 0.33695652, 0.32608696, 0.33695652,
       0.33695652, 0.32608696, 0.33695652, 0.33695652, 0.32608696,
       0.33695652, 0.33695652, 0.32608696, 0.33695652, 0.33695652,
       0.32608696, 0.33695652, 0.33695652, 0.32608696, 0.33695652,
       0.33695652, 0.32608696, 0.33695652, 0.33695652, 0.32608696,
       0.33695652, 0.33695652, 0.32608696, 0.33695652, 0.33695652,
       0.32608696, 0.33695652, 0.33695652, 0.32608696, 0.33695652,
       0.33695652, 0.32608696, 0.33695652, 0.33695652, 0.32608696,
       0.33695652, 0.33695652, 0.32608696, 0.33695652, 0.33695652,
       0.32608696, 0.33695652, 0.33695652, 0.32608696, 0.33695652,
       0.33695652, 0.32608696, 0.33695652, 0.33695652, 0.32608696,
       0.33695652, 0.33695652, 0.32608696, 0.33695652, 0.33695652,
       0.32608696, 0.33695652])
Coordinates:
  * time     (time) datetime64[ns] 1982-03-01 1982-04-01 ... 2020-05-01
    year     (time) int64 1982 1982 1982 1983 1983 ... 2019 2019 2020 2020 2020
```

```
[16]: # Test that the sum of the weights for the season is 1.0
np.testing.assert_allclose(weights.groupby('time.year').sum().values, np.ones(39))  ##␣
↪the weight is one for each year
```

```
[30]: # Calculate the weighted average
SEAS5_Siberia_weighted = (SEAS5_Siberia * weights).groupby('time.year').sum(dim='time
↪', min_count = 3)
SEAS5_Siberia_weighted
```

```
[30]: <xarray.Dataset>
Dimensions:    (latitude: 41, leadtime: 3, longitude: 132, number: 51, year: 39)
```

(continues on next page)

(continued from previous page)

```
Coordinates:
  * leadtime    (leadtime) int64 2 3 4
  * latitude    (latitude) float32 70.0 69.0 68.0 67.0 ... 33.0 32.0 31.0 30.0
  * number      (number) int64 0 1 2 3 4 5 6 7 8 9 ... 42 43 44 45 46 47 48 49 50
  * longitude   (longitude) float32 -11.0 -10.0 -9.0 -8.0 ... 118.0 119.0 120.0
  * year        (year) int64 1982 1983 1984 1985 1986 ... 2017 2018 2019 2020
Data variables:
    t2m         (year, leadtime, number, latitude, longitude) float64 268.8 ... 289.9
    d2m         (year, leadtime, number, latitude, longitude) float64 266.1 ... 286.0
```

Or as function:

[31]:
```
def season_mean(ds, years, calendar='standard'):
    # Make a DataArray with the number of days in each month, size = len(time)
    month_length = ds.time.dt.days_in_month

    # Calculate the weights by grouping by 'time.season'
    weights = month_length.groupby('time.year') / month_length.groupby('time.year').
↪sum()

    # Test that the sum of the weights for each season is 1.0
    np.testing.assert_allclose(weights.groupby('time.year').sum().values, np.
↪ones(years))

    # Calculate the weighted average
    return (ds * weights).groupby('time.year').sum(dim='time', min_count = 3)
```

[33]:
```
ERA5_Siberia_weighted = season_mean(ERA5_Siberia, years = 42)
ERA5_Siberia_weighted
```

[33]:
```
<xarray.Dataset>
Dimensions:    (latitude: 41, longitude: 132, year: 42)
Coordinates:
  * latitude   (latitude) float32 70.0 69.0 68.0 67.0 ... 33.0 32.0 31.0 30.0
  * longitude  (longitude) float32 -11.0 -10.0 -9.0 -8.0 ... 118.0 119.0 120.0
  * year       (year) int64 1979 1980 1981 1982 1983 ... 2017 2018 2019 2020
Data variables:
    t2m        (year, latitude, longitude) float64 270.5 270.9 ... 289.1 290.6
    d2m        (year, latitude, longitude) float64 267.1 267.4 ... 283.5 284.7
```

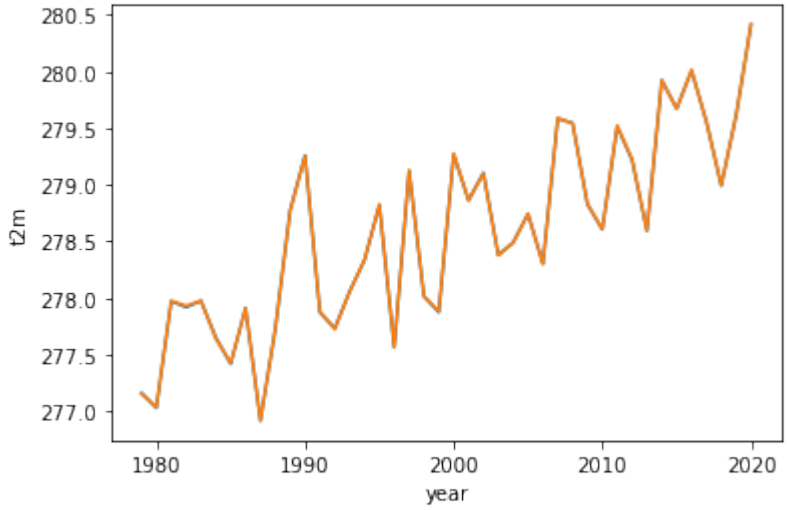
What is the difference between the mean and weighted mean?

Barely visible the difference

[34]:
```
ERA5_Siberia_weighted['t2m'].mean(['longitude', 'latitude']).plot()
ERA5_Siberia['t2m'].groupby('time.year').mean().mean(['longitude','latitude']).plot()
```

[34]:
```
[<matplotlib.lines.Line2D at 0x2aedbb30ca90>]
```

[34]:
```
[<matplotlib.lines.Line2D at 0x2aed3bf98b50>]
```



Spatial selection

What spatial extent defines the event you are analyzing? The easiest option is to select a lat-lon box, like we did for the Siberian heatwave example (i.e we average the temperature over 30-70N, -11-120E).

In case you want to specify another domain than a lat-lon box, you could mask the datasets. For the California Fires example, we select the domain with high temperature anomalies (>2 standard deviation), see California_august_temperature_anomaly. For the UK example, we want a country-averaged timeseries instead of a box. In this case, we use another observational product: the EOBS dataset that covers Europe. We *upscale* this dataset to the same resolution as SEAS5 and create a *mask* to take the spatial average over the UK, see *Using EOBS + upscaling*.

[35]:
```
ERA5_Siberia_events = (
    ERA5_Siberia_weighted['t2m'].sel(  # Select 2 metre temperature
        latitude=slice(70, 30),        # Select the latitudes
        longitude=slice(-11, 120)).    # Select the longitude
    mean(['longitude', 'latitude']))   # And average
ERA5_Siberia_events
```

[35]:
```
<xarray.DataArray 't2m' (year: 42)>
array([277.16074991, 277.03096514, 277.97289103, 277.91908656,
       277.97520289, 277.65083809, 277.41882817, 277.91126484,
       276.91748994, 277.7290794 , 278.77625033, 279.26028161,
       277.8734671 , 277.73297789, 278.05803743, 278.33474762,
       278.8209804 , 277.56676759, 279.12484889, 278.01516085,
       277.87531784, 279.26812975, 278.86348174, 279.10661042,
       278.37979554, 278.49025936, 278.74264131, 278.30305153,
       279.58431666, 279.54500525, 278.82677061, 278.60606811,
       279.51421559, 279.21937005, 278.59269493, 279.92328752,
       279.6721145 , 280.00902662, 279.55917477, 278.99138847,
       279.61443777, 280.41497365])
Coordinates:
  * year     (year) int64 1979 1980 1981 1982 1983 ... 2016 2017 2018 2019 2020
```

In addition to the large domain, we select one more specific domain that faced the highest anomalies, also used here

[36]:
```
ERA5_Siberia_events_zoomed = (
    ERA5_Siberia_weighted['t2m'].sel(  # Select 2 metre temperature
```

(continues on next page)

(continued from previous page)

```
        latitude=slice(70, 50),         # Select the latitudes
        longitude=slice(65, 120)).     # Select the longitude
    mean(['longitude', 'latitude']))
```

And we repeat this for the SEAS5 events

[37]:
```
SEAS5_Siberia_events = (
    SEAS5_Siberia_weighted['t2m'].sel(
        latitude=slice(70, 30),
        longitude=slice(-11, 120)).
    mean(['longitude', 'latitude']))
SEAS5_Siberia_events.load()
```

```
/soge-home/users/cenv0732/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/xarray/
↪core/nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

[37]:
```
<xarray.DataArray 't2m' (year: 39, leadtime: 3, number: 51)>
array([[[277.70246026, 277.08011538, 277.05805243, ...,          nan,
                 nan,          nan],
        [276.54063304, 277.63527276, 276.53540684, ...,          nan,
                 nan,          nan],
        [276.94382457, 277.22540106, 277.25375804, ...,          nan,
                 nan,          nan]],

       [[276.68638666, 276.64418409, 276.88169219, ...,          nan,
                 nan,          nan],
        [277.06362955, 277.30470221, 276.49967939, ...,          nan,
                 nan,          nan],
        [276.37166345, 276.63563118, 277.05456392, ...,          nan,
                 nan,          nan]],

       [[277.53103277, 277.49691758, 278.32115366, ...,          nan,
                 nan,          nan],
        [277.53911427, 278.11393678, 277.66278741, ...,          nan,
                 nan,          nan],
        [278.24589375, 277.71341079, 277.3067712 , ...,          nan,
                 nan,          nan]],

       ...,

       [[278.78906418, 278.0626699 , 278.09438675, ..., 278.10947691,
         278.27620377, 278.18620179],
        [278.65929139, 277.33954004, 278.65951576, ..., 278.22959696,
         278.32163068, 278.94724957],
        [278.90689211, 277.9030209 , 279.13818072, ..., 278.76767259,
         279.13397914, 277.76992423]],

       [[278.6218426 , 277.95006232, 278.22900254, ..., 277.56330007,
         277.99480916, 277.66857676],
        [278.39808792, 277.65889255, 277.92266928, ..., 278.39390445,
         277.90353039, 278.01793147],
        [278.63429219, 278.11630486, 278.58465727, ..., 277.7081865 ,
         277.73949614, 278.65482764]],

       [[279.28124227, 278.53474142, 278.47436518, ..., 278.93209185,
         278.11716261, 279.3904762 ],
        [277.77935773, 279.15571385, 279.02168652, ..., 279.25803913,
```

(continues on next page)

(continued from previous page)

```
         278.8991169 , 278.72803803],
        [278.54721722, 278.25816177, 279.65502139, ..., 279.01242149,
         278.80174459, 278.23615329]]])
Coordinates:
  * leadtime  (leadtime) int64 2 3 4
  * number    (number) int64 0 1 2 3 4 5 6 7 8 9 ... 42 43 44 45 46 47 48 49 50
  * year      (year) int64 1982 1983 1984 1985 1986 ... 2016 2017 2018 2019 2020
```

[39]:
```
SEAS5_Siberia_events_zoomed = (
    SEAS5_Siberia_weighted['t2m'].sel(
        latitude=slice(70, 50),
        longitude=slice(65, 120)).
    mean(['longitude', 'latitude']))
SEAS5_Siberia_events_zoomed.load()
```

[39]:
```
<xarray.DataArray 't2m' (year: 39, leadtime: 3, number: 51)>
array([[[269.41349502, 267.46724112, 268.92858923, ...,          nan,
                 nan,          nan],
        [267.30541714, 269.45786213, 267.88083134, ...,          nan,
                 nan,          nan],
        [266.70463988, 269.36559057, 267.85380896, ...,          nan,
                 nan,          nan]],

       [[267.65255078, 267.85459971, 267.93397041, ...,          nan,
                 nan,          nan],
        [267.86908721, 269.16875401, 266.25505375, ...,          nan,
                 nan,          nan],
        [267.37705396, 268.0216673 , 269.216725  , ...,          nan,
                 nan,          nan]],

       [[269.11559244, 270.30993792, 268.97992022, ...,          nan,
                 nan,          nan],
        [268.42632866, 270.23730451, 268.14872887, ...,          nan,
                 nan,          nan],
        [269.91675564, 269.37623754, 268.62430776, ...,          nan,
                 nan,          nan]],

       ...,

       [[270.1395106 , 269.2763681 , 269.63408345, ..., 267.98167445,
         270.99148119, 269.2334804 ],
        [269.47838693, 267.62436995, 270.46814617, ..., 269.60061317,
         269.2030878 , 270.34174847],
        [271.39762301, 268.06102893, 271.27933216, ..., 269.56866515,
         270.64943148, 266.36754614]],

       [[269.32829344, 268.54758963, 269.10385302, ..., 268.03762412,
         269.16690743, 268.41760566],
        [269.09579833, 268.27844902, 268.80430384, ..., 269.38437353,
         269.58112924, 269.15390309],
        [269.2755198 , 269.28402279, 270.34429505, ..., 268.05044769,
         269.01128231, 270.09822017]],

       [[271.35447849, 269.73507649, 269.16852164, ..., 270.77523033,
         268.30803885, 271.87237937],
        [269.33923167, 270.81122764, 269.97102764, ..., 272.56708782,
         270.69195734, 269.90622653],
```

(continues on next page)

(continued from previous page)

```
        [270.37695877, 269.72247595, 272.524012  , ..., 270.04026793,
         269.81585811, 267.24061429]]])
Coordinates:
  * leadtime  (leadtime) int64 2 3 4
  * number    (number) int64 0 1 2 3 4 5 6 7 8 9 ... 42 43 44 45 46 47 48 49 50
  * year      (year) int64 1982 1983 1984 1985 1986 ... 2016 2017 2018 2019 2020
```

[40]:
```
SEAS5_Siberia_events.to_dataframe().to_csv('Data/SEAS5_Siberia_events.csv')
ERA5_Siberia_events.to_dataframe().to_csv('Data/ERA5_Siberia_events.csv')
```

[41]:
```
SEAS5_Siberia_events_zoomed.to_dataframe().to_csv('Data/SEAS5_Siberia_events_zoomed.
↪csv')
ERA5_Siberia_events_zoomed.to_dataframe().to_csv('Data/ERA5_Siberia_events_zoomed.csv
↪')
```

1.6 Evaluate

Can seasonal forecasts be used as 'alternate' realities? Here we show how a set of evaluation metrics can be used to answer this question. The evaluation metrics are available through an R package for easy evaluation of the UNSEEN ensemble. Here, we illustrate how this package can be used in the UNSEEN workflow. We will evaluate the generated UNSEEN ensemble of UK February precipitation and of MAM Siberian heatwaves.

The framework to evaluate the UNSEEN ensemble presented here consists of testing the ensemble member independence, model stability and model fidelity, see also NPJ preprint.

Note

This is R code and not python!

We switch to R since we believe R has a better functionality in extreme value statistics.

We load the UNSEEN package and read in the data.

[2]:
```
library(UNSEEN)
```

The data that is imported here are the files stored at the end of the *preprocessing step*.

[125]:
```
SEAS5_Siberia_events <- read.csv("Data/SEAS5_Siberia_events.csv",␣
↪stringsAsFactors=FALSE)
ERA5_Siberia_events <- read.csv("Data/ERA5_Siberia_events.csv",␣
↪stringsAsFactors=FALSE)
```

[126]:
```
SEAS5_Siberia_events_zoomed <- read.csv("Data/SEAS5_Siberia_events_zoomed.csv",␣
↪stringsAsFactors=FALSE)
ERA5_Siberia_events_zoomed <- read.csv("Data/ERA5_Siberia_events_zoomed.csv",␣
↪stringsAsFactors=FALSE)
```

[127]:
```
SEAS5_Siberia_events$t2m <- SEAS5_Siberia_events$t2m - 273.15
ERA5_Siberia_events$t2m <- ERA5_Siberia_events$t2m - 273.15
SEAS5_Siberia_events_zoomed$t2m <- SEAS5_Siberia_events_zoomed$t2m - 273.15
ERA5_Siberia_events_zoomed$t2m <- ERA5_Siberia_events_zoomed$t2m - 273.15
```

[6]:
```
head(SEAS5_Siberia_events_zoomed,n = 3)
head(ERA5_Siberia_events, n = 3)
```

A data.frame: 3 × 4

	year <int>	leadtime <int>	number <int>	t2m <dbl>
1	1982	2	0	-3.736505
2	1982	2	1	-5.682759
3	1982	2	2	-4.221411

A data.frame: 3 × 2

	year <int>	t2m <dbl>
1	1979	4.010750
2	1980	3.880965
3	1981	4.822891

[7]:
```
EOBS_UK_weighted_df <- read.csv("Data/EOBS_UK_weighted_upscaled.csv",␣
↪stringsAsFactors=FALSE)
SEAS5_UK_weighted_df <- read.csv("Data/SEAS5_UK_weighted_masked.csv",␣
↪stringsAsFactors=FALSE)
```

And then convert the time class to Date format, with the ymd function in lubridate:

[42]:
```
EOBS_UK_weighted_df$time <- lubridate::ymd(EOBS_UK_weighted_df$time)
str(EOBS_UK_weighted_df)

EOBS_UK_weighted_df_hindcast <- EOBS_UK_weighted_df[
    EOBS_UK_weighted_df$time > '1982-02-01' &
    EOBS_UK_weighted_df$time < '2017-02-01',
    ]


SEAS5_UK_weighted_df$time <- lubridate::ymd(SEAS5_UK_weighted_df$time)
str(SEAS5_UK_weighted_df)
```

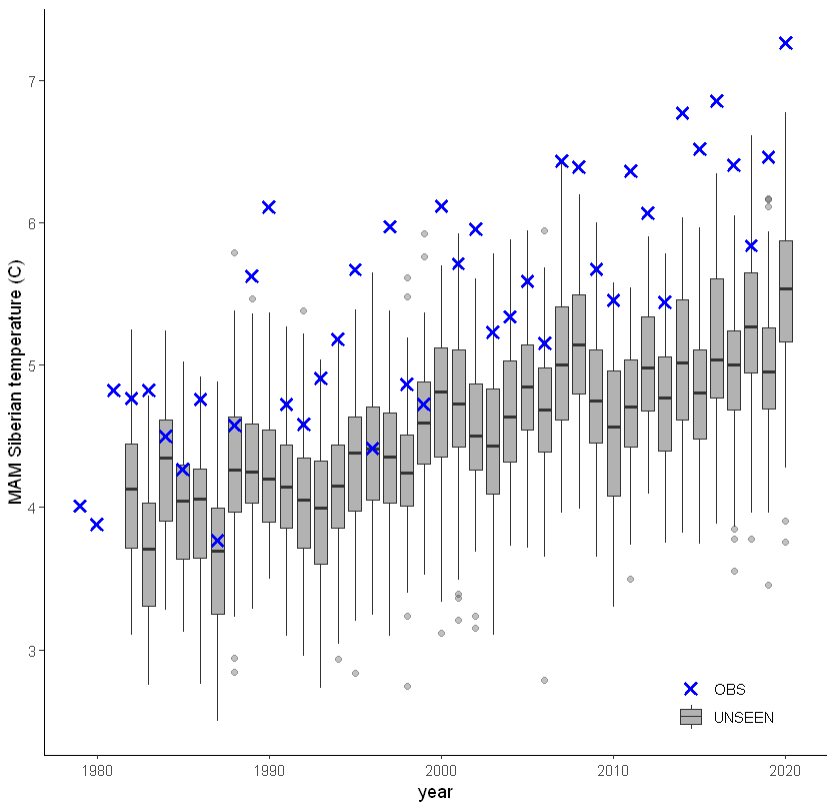
```
'data.frame':   71 obs. of  2 variables:
 $ time: Date, format: "1950-02-28" "1951-02-28" ...
 $ rr  : num  4.13 3.25 1.07 1.59 2.59 ...
'data.frame':   4375 obs. of  4 variables:
 $ leadtime: int  2 2 2 2 2 2 2 2 2 2 ...
 $ number  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ time    : Date, format: "1982-02-01" "1983-02-01" ...
 $ tprate  : num  1.62 2.93 3.27 2 3.31 ...
```

1.6.1 Timeseries

Here we plot the timeseries of SEAS5 (UNSEEN) and ERA5 (OBS) for the entire domain and a zoomed domain for the Siberian Heatwave.

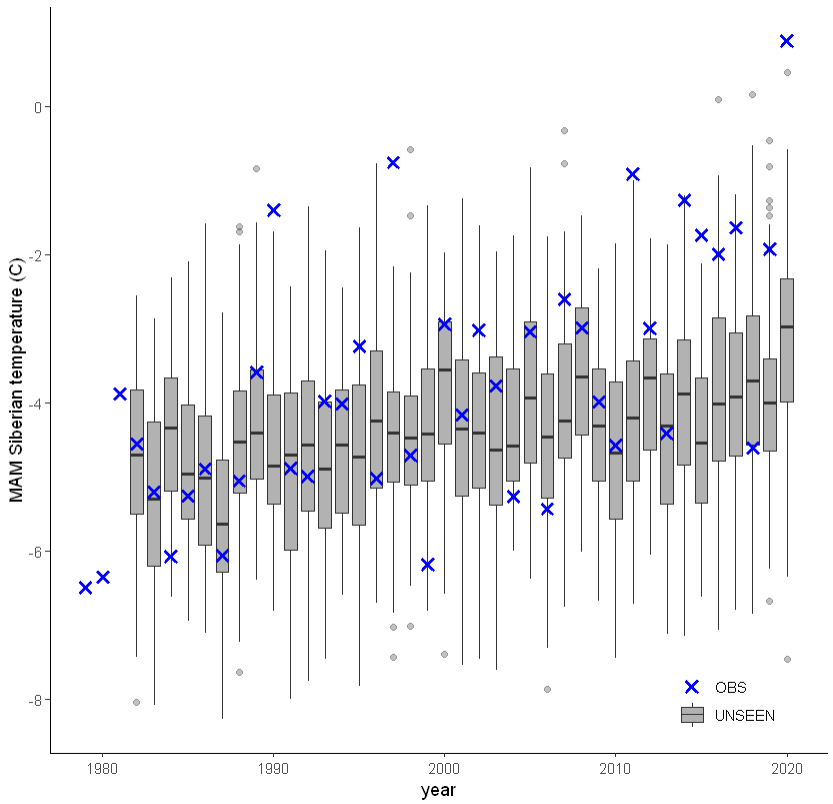
[9]:
```
unseen_timeseries(
    ensemble = SEAS5_Siberia_events,
    obs = ERA5_Siberia_events,
    ensemble_yname = "t2m",
    ensemble_xname = "year",
    obs_yname = "t2m",
    obs_xname = "year",
    ylab = "MAM Siberian temperature (C)")
```

```
Warning message:
"Removed 2756 rows containing non-finite values (stat_boxplot)."
```



[10]:
```
unseen_timeseries(
    ensemble = SEAS5_Siberia_events_zoomed,
    obs = ERA5_Siberia_events_zoomed,
    ensemble_yname = "t2m",
    ensemble_xname = "year",
    obs_yname = "t2m",
    obs_xname = "year",
    ylab = "MAM Siberian temperature (C)")
```

```
Warning message:
"Removed 2756 rows containing non-finite values (stat_boxplot)."
```



This shows that there is a temperature trend over the entire domain. Here we will continue with the 'zoomed' domain because it better describes the 2020 event.

The timeseries consist of hindcast (years 1982-2016) and archived forecasts (years 2017-2020). The datasets are slightly different: the hindcasts contains 25 members whereas operational forecasts contain 51 members, the native resolution is different and the dataset from which the forecasts are initialized is different.

For the evaluation of the UNSEEN ensemble we want to only use the SEAS5 hindcasts for a consistent dataset. Note, 2017 is not used in either the hindcast nor the operational dataset in this example, since it contains forecasts both initialized in 2016 (hindcast) and 2017 (forecast), see *retrieve*. We split SEAS5 into hindcast and operational forecasts:

[24]:
```
SEAS5_Siberia_events_zoomed_hindcast <- SEAS5_Siberia_events_zoomed[
    SEAS5_Siberia_events_zoomed$year < 2017 &
    SEAS5_Siberia_events_zoomed$number < 25,]

SEAS5_Siberia_events_zoomed_forecasts <- SEAS5_Siberia_events_zoomed[
    SEAS5_Siberia_events_zoomed$year > 2017,]
```

And we select the same years for ERA5.

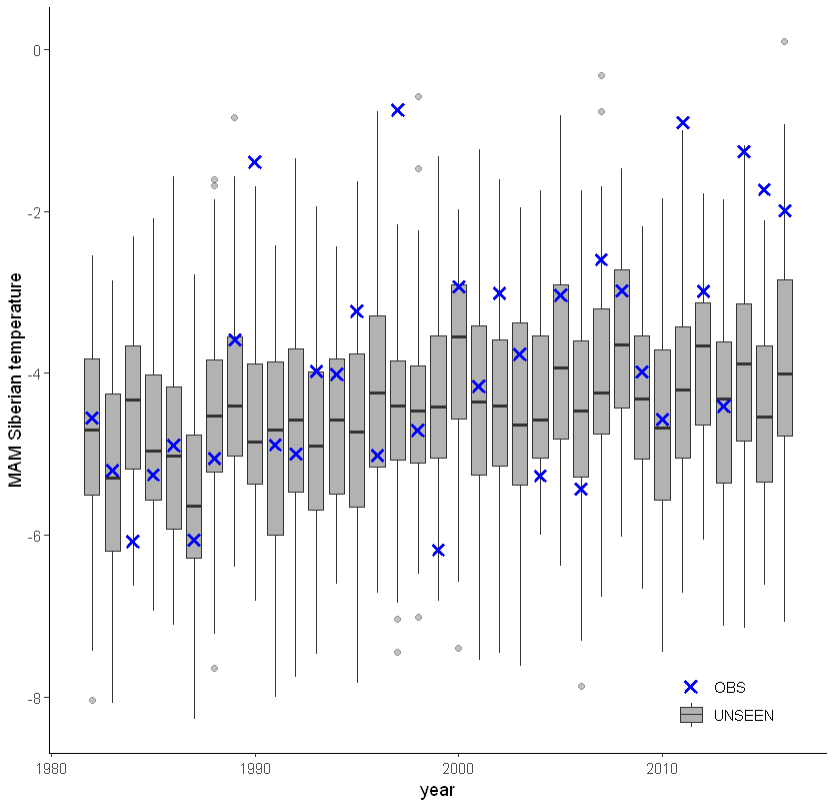
[32]:
```
ERA5_Siberia_events_zoomed_hindcast <- ERA5_Siberia_events_zoomed[
    ERA5_Siberia_events_zoomed$year < 2017 &
    ERA5_Siberia_events_zoomed$year > 1981,]
```

[33]:
```
unseen_timeseries(
```

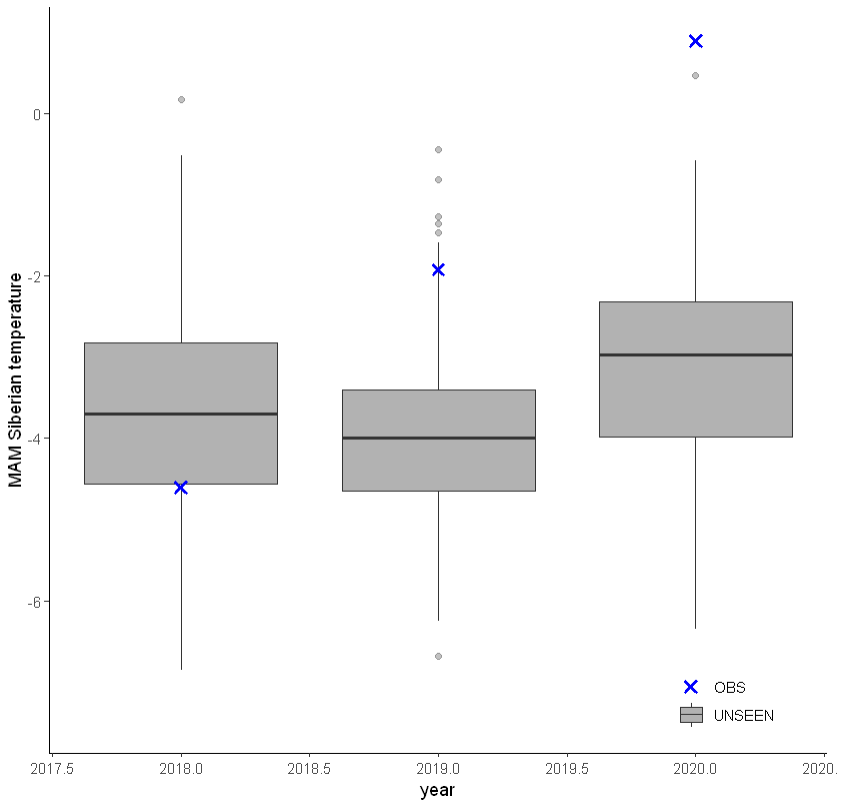
(continues on next page)

(continued from previous page)

```
    ensemble = SEAS5_Siberia_events_zoomed_hindcast,
    obs = ERA5_Siberia_events_zoomed_hindcast,
    ensemble_yname = "t2m",
    ensemble_xname = "year",
    obs_yname = "t2m",
    obs_xname = "year",
    ylab = "MAM Siberian temperature")
```

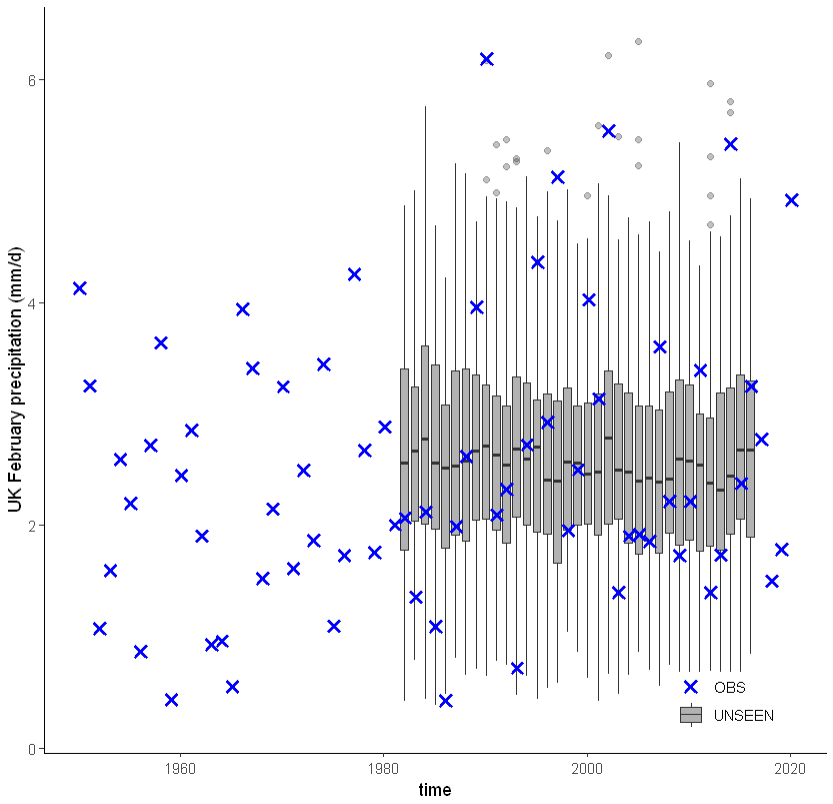


[26]:
```
unseen_timeseries(
    ensemble = SEAS5_Siberia_events_zoomed_forecasts,
    obs = ERA5_Siberia_events_zoomed[ERA5_Siberia_events_zoomed$year > 2017,],
    ensemble_yname = "t2m",
    ensemble_xname = "year",
    obs_yname = "t2m",
    obs_xname = "year",
    ylab = "MAM Siberian temperature")
```

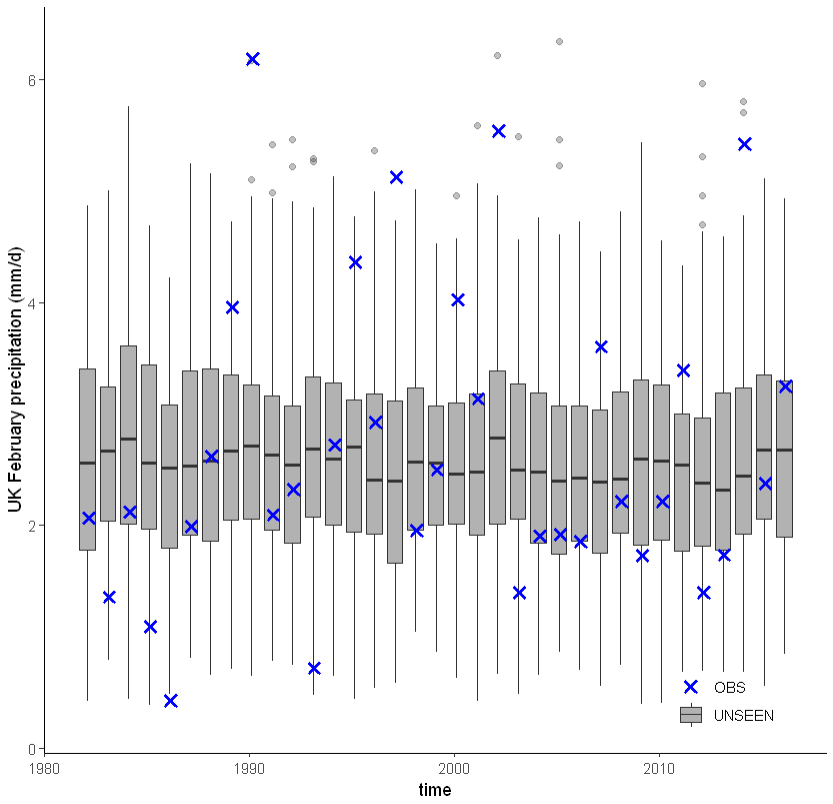


For the UK we have a longer historical record available from EOBS:

[12]:
```
unseen_timeseries(ensemble = SEAS5_UK_weighted_df,
                  obs = EOBS_UK_weighted_df,
                  ylab = 'UK February precipitation (mm/d)')
```



[38]:
```
unseen_timeseries(ensemble = SEAS5_UK_weighted_df,
                  obs = EOBS_UK_weighted_df_hindcast,
                  ylab = 'UK February precipitation (mm/d)')
```



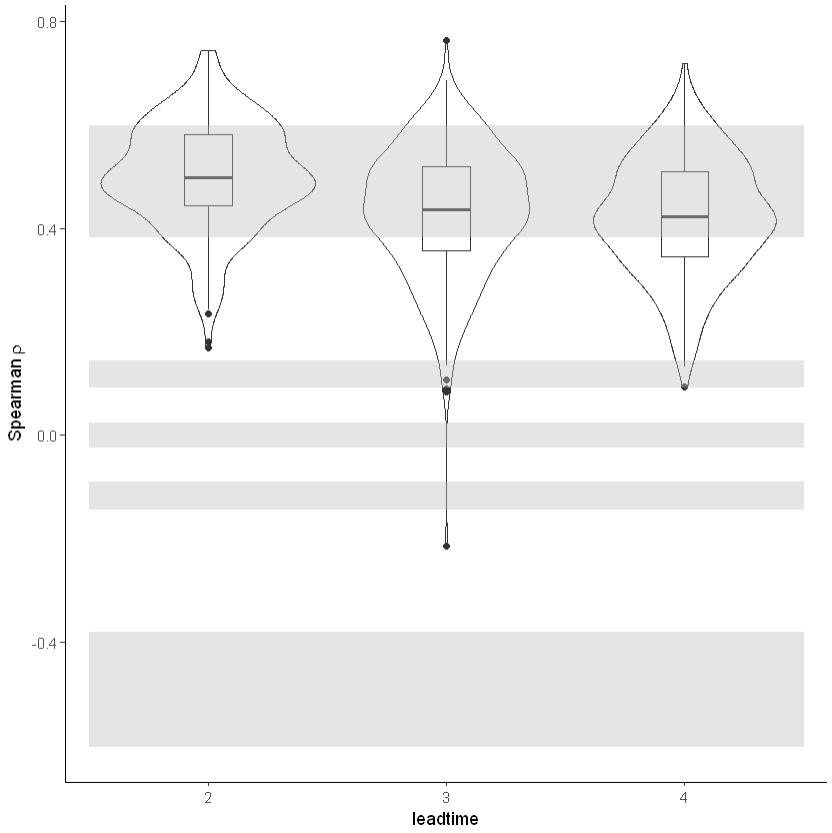
Call the documentation of the function with `?unseen_timeseries`

1.6.2 Independence

Significance ranges need fixing + detrend method (Rob)

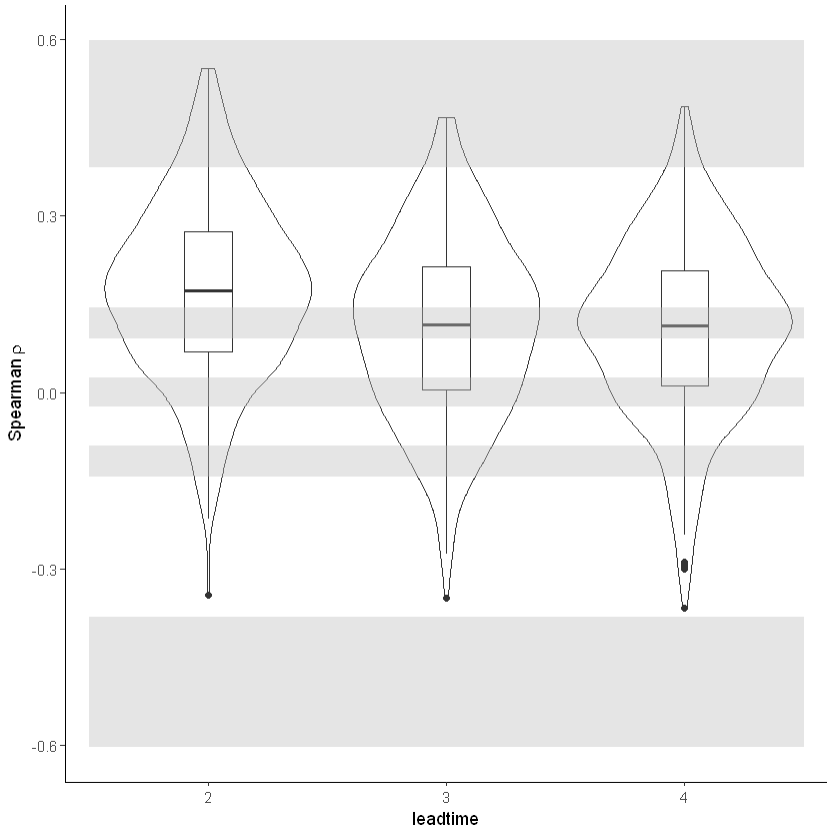
[14]:
```
independence_test(
    ensemble = SEAS5_Siberia_events,
    n_lds = 3,
    var_name = "t2m",
)
```

```
Warning message:
"Removed 975 rows containing non-finite values (stat_ydensity)."
Warning message:
"Removed 975 rows containing non-finite values (stat_boxplot)."
```



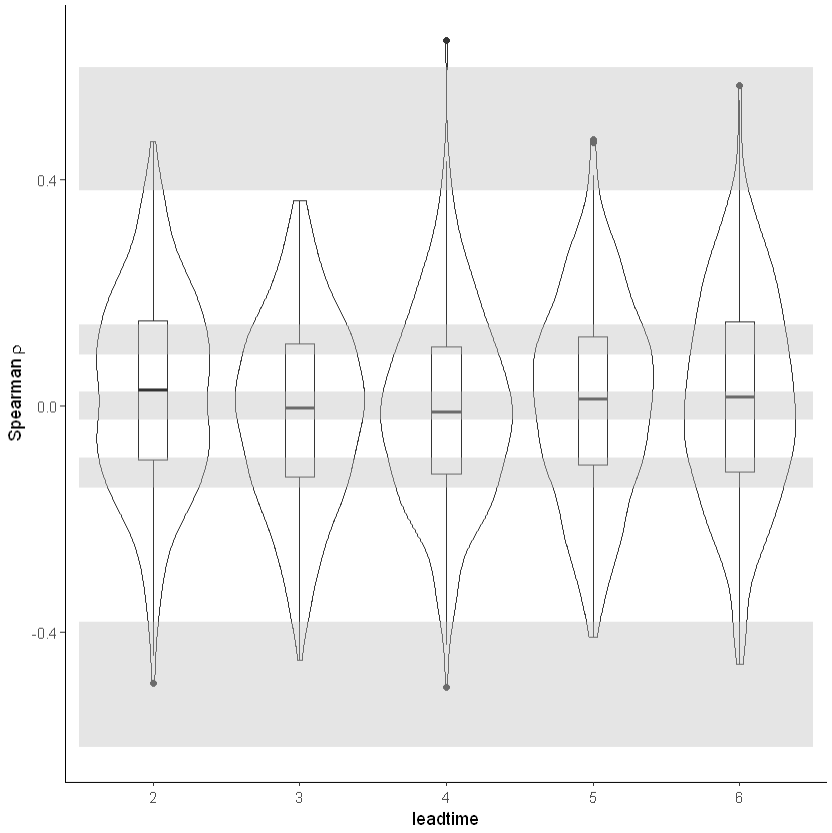
[15]:
```
independence_test(
    ensemble = SEAS5_Siberia_events_zoomed,
    n_lds = 3,
    var_name = "t2m",
)
```

```
Warning message:
"Removed 975 rows containing non-finite values (stat_ydensity)."
Warning message:
"Removed 975 rows containing non-finite values (stat_boxplot)."
```



[16]:
```
independence_test(ensemble = SEAS5_UK)
```

```
Warning message:
"Removed 1625 rows containing non-finite values (stat_ydensity)."
Warning message:
"Removed 1625 rows containing non-finite values (stat_boxplot)."
```

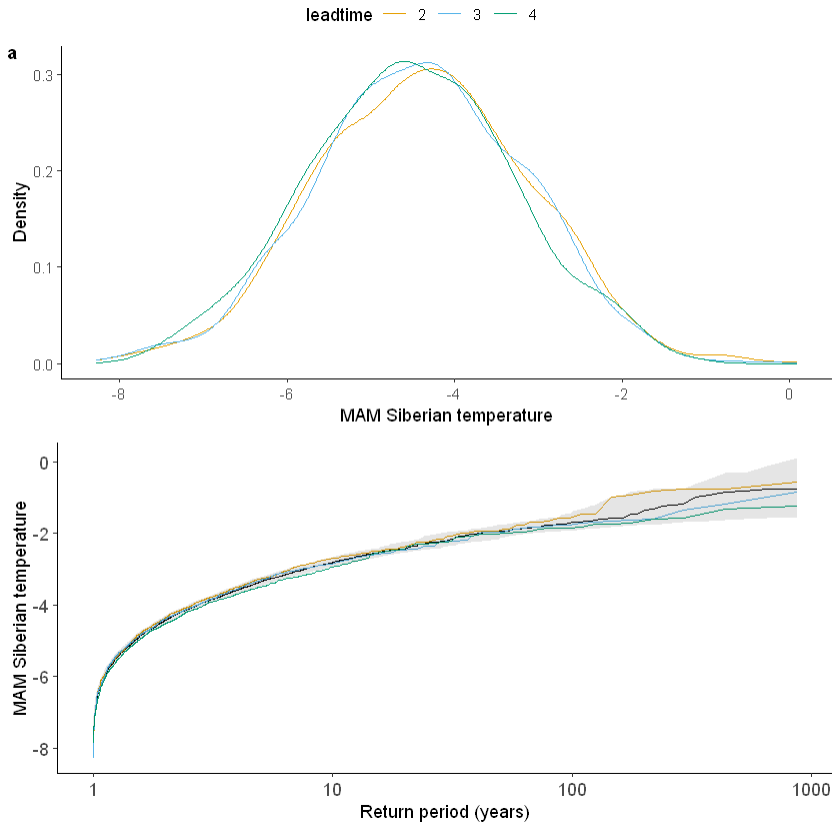


1.6.3 Stability

For the stability test we assess whether the events get more severe with leadtime, due to a potential 'drift' in the model. We need to use the consistent hindcast dataset for this.

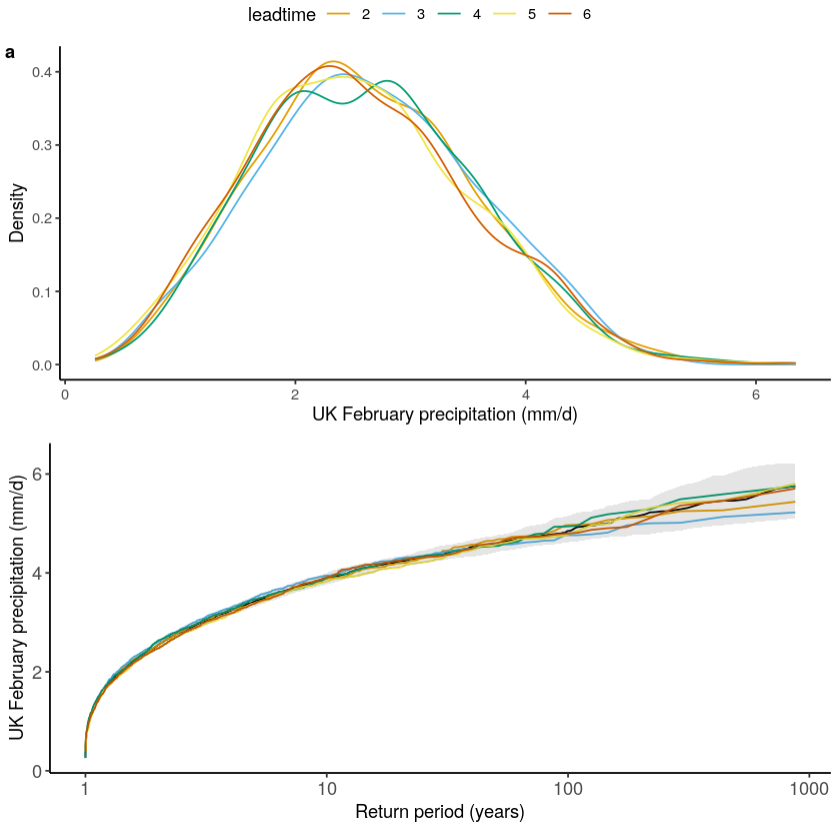
[27]:
```
stability_test(
    ensemble = SEAS5_Siberia_events_zoomed_hindcast,
    lab = 'MAM Siberian temperature',
    var_name = 't2m'
)
```

```
Warning message:
"Removed 2 row(s) containing missing values (geom_path)."
```



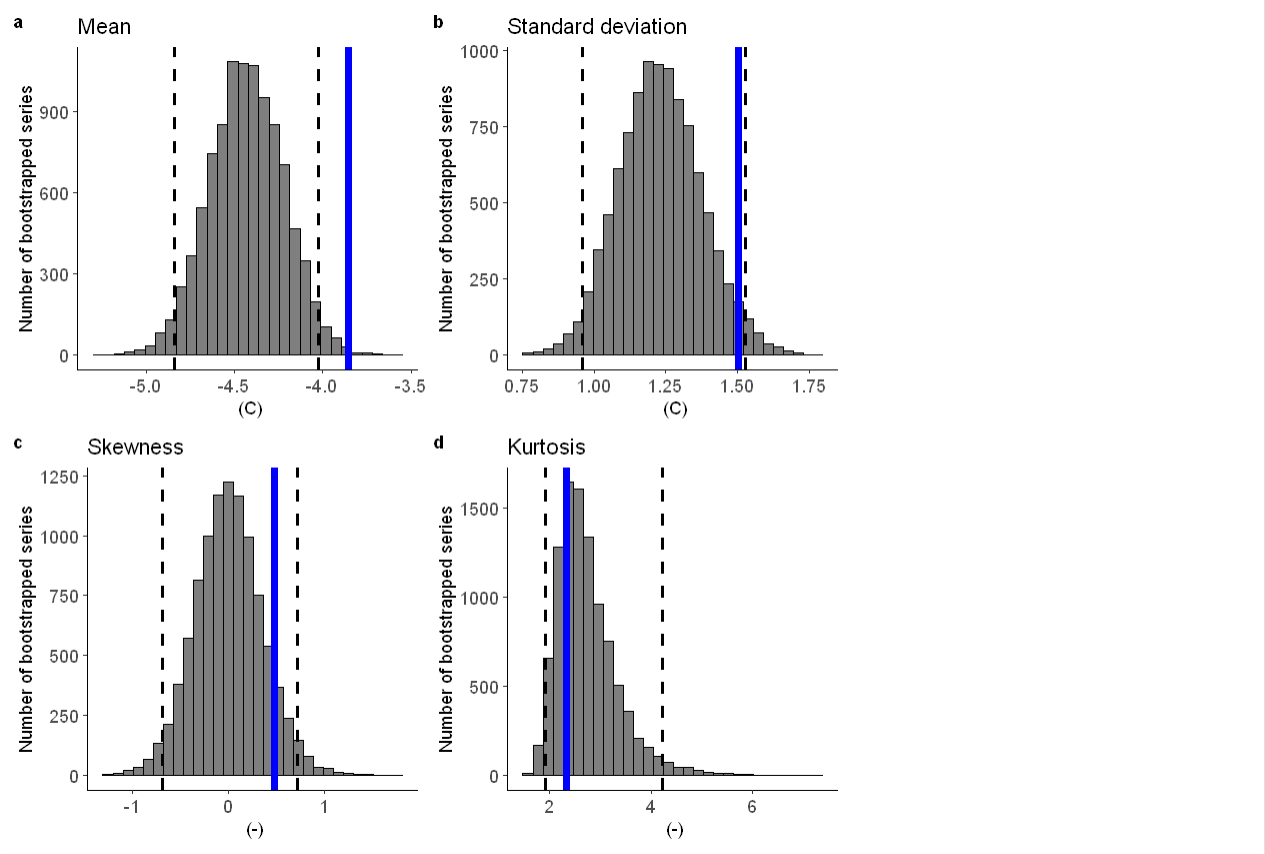
[8]:
```
stability_test(ensemble = SEAS5_UK, lab = 'UK February precipitation (mm/d)')
```

```
Warning message:
"Removed 4 row(s) containing missing values (geom_path)."
```



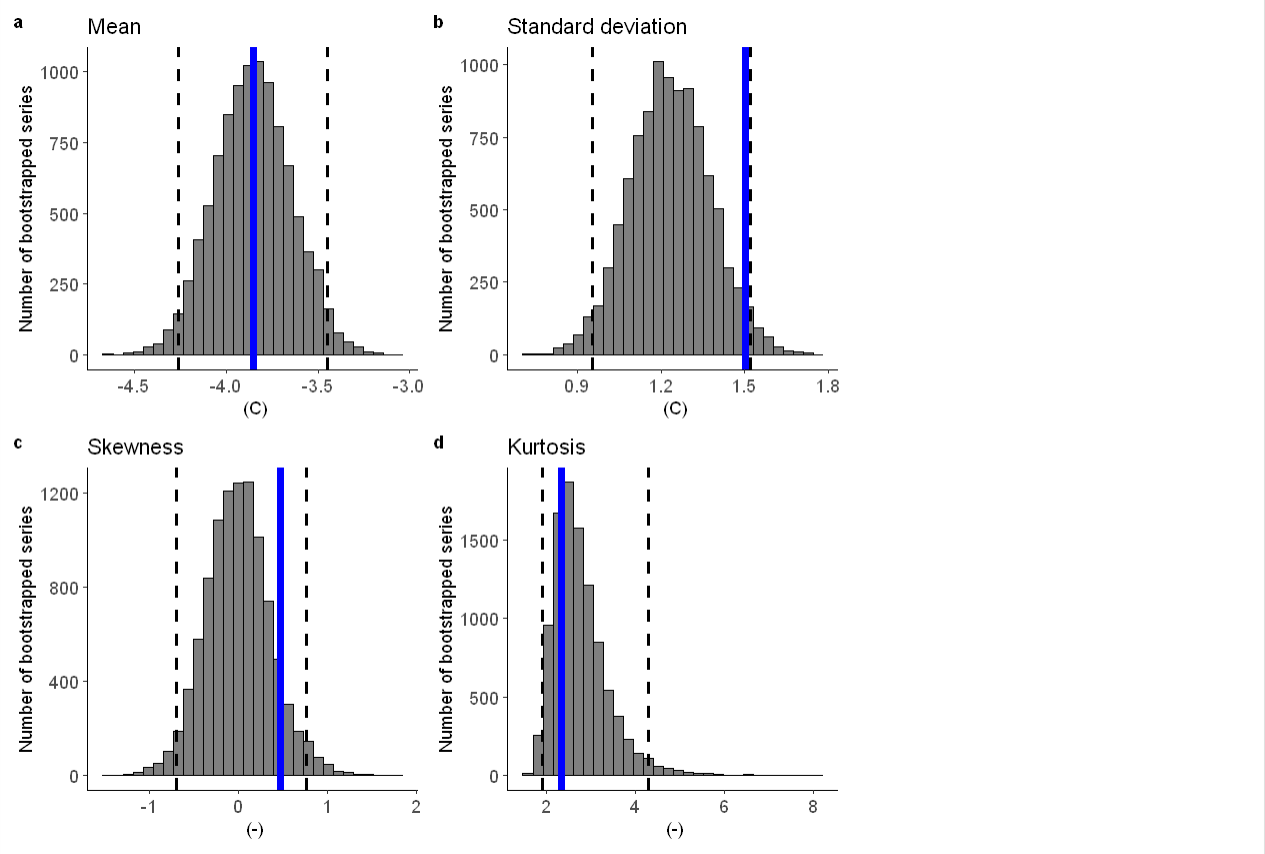
1.6.4 Fidelity

[92]:
```
fidelity_test(
    obs = ERA5_Siberia_events_zoomed_hindcast$t2m,
    ensemble = SEAS5_Siberia_events_zoomed_hindcast$t2m,
    units = 'C',
    biascor = FALSE
)
```

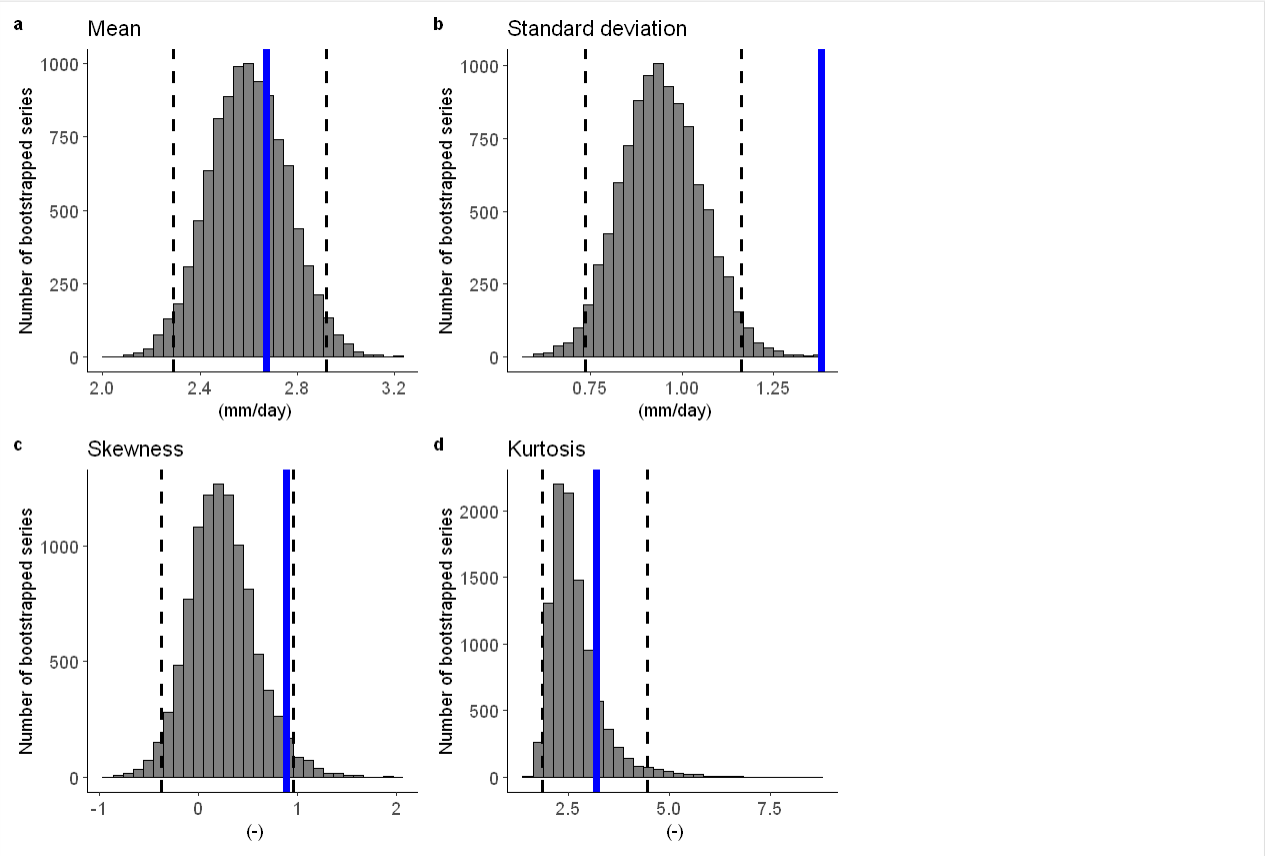


Lets apply a additive biascor

[90]:
```
#Lets apply a additive biascor
obs = ERA5_Siberia_events_zoomed_hindcast$t2m
ensemble = SEAS5_Siberia_events_zoomed_hindcast$t2m
ensemble_biascor = ensemble + (mean(obs) - mean(ensemble))

fidelity_test(
    obs = obs,
    ensemble = ensemble_biascor,
    units = 'C',
    biascor = FALSE
)
```

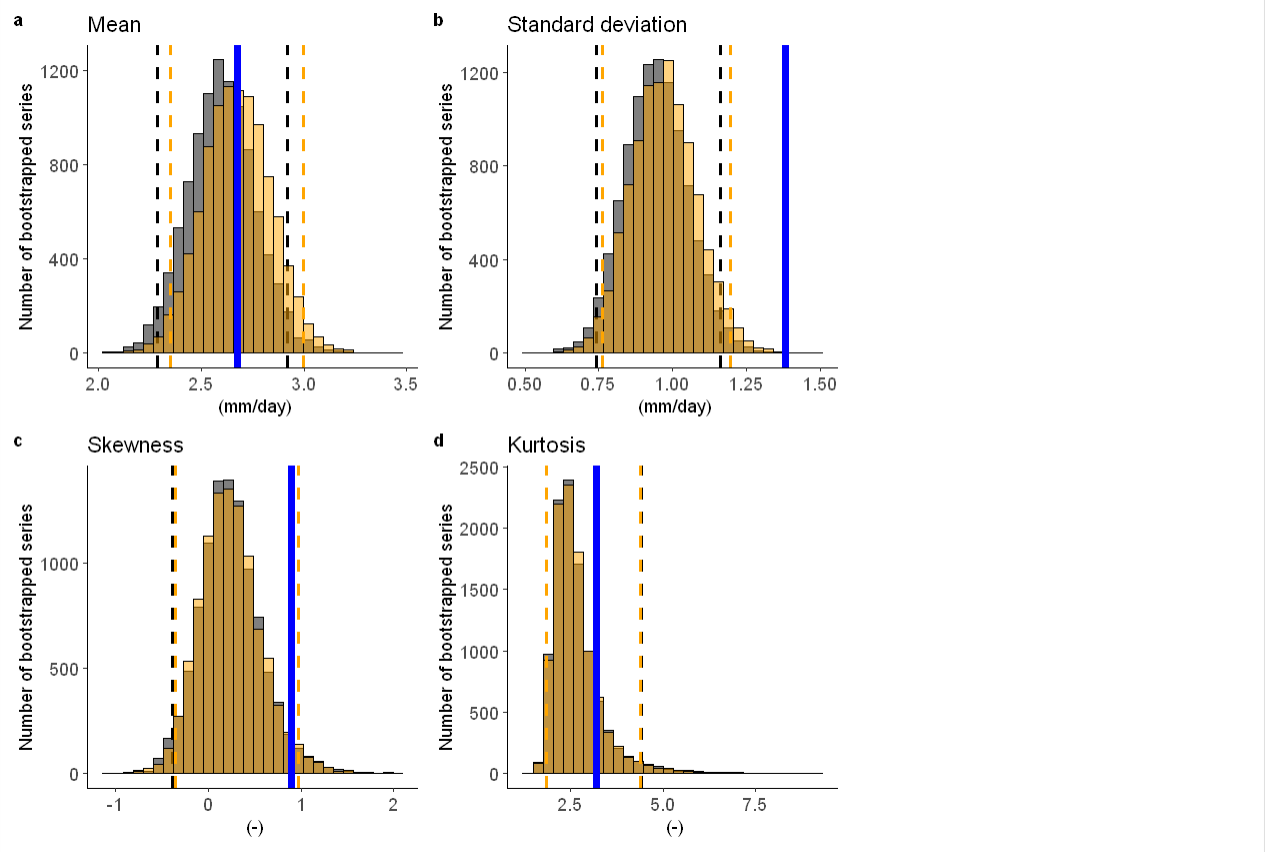


[47]:
```
fidelity_test(obs = EOBS_UK_weighted_df_hindcast$rr, ensemble = SEAS5_UK_weighted_df
↪$tprate)
```



To include a mean-bias correction, set `biascor = TRUE`:

[46]:
```
fidelity_test(obs = EOBS_UK_weighted_df_hindcast$rr, ensemble = SEAS5_UK_weighted_df
↪$tprate, biascor = TRUE)
```



[16]:
```
?fidelity_test
```

1.7 Illustrate

Here we use extreme value theory (EVT) to fit extreme value distributions to the SEAS5 (UNSEEN) and ERA5 (observed) data.

To see example applications, have a look at the examples:

- Siberian Heatwave
- California fires
- UK Precipitation

We define a function to plot the extreme value distributions:

[75]:
```
library(extRemes)
library(ggplot2)
library(ggpubr)

EVT_plot <- function(obs, ensemble, GEV_type, main, y_lab = "February average␣
↪precipitation (mm/day)", ylim = NA) {
  ## We plot the GEV distribution for ERA5 and empirical data for SEAS5
  fit_obs <- fevd(
    x = obs, threshold = NULL, threshold.fun = ~1, location.fun = ~1,
```

(continues on next page)

(continued from previous page)

```
    scale.fun = ~1, shape.fun = ~1, use.phi = FALSE,
    type = GEV_type, method = "MLE", initial = NULL, # type= c("GEV", "GP", "PP",
↪"Gumbel", "Exponential"), method= c("MLE", "GMLE", "Bayesian", "Lmoments")
    span = NULL, units = NULL, time.units = "days", period.basis = "year", ## time␣
↪and period only important for labelling and do not influence the calculation
    na.action = na.fail, optim.args = NULL, priorFun = NULL,
    priorParams = NULL, proposalFun = NULL, proposalParams = NULL,
    iter = 9999, weights = 1, blocks = NULL, verbose = FALSE
  )

  ## Now calculate the return levels and their confidence intervals for each return␣
↪period within rperiods
  rperiods <- c(seq(from = 1.01, to = 1.5, by = 0.1), 1.7, 2, 3, 5, 10, 20, 50, 80,␣
↪100, 120, 200, 250, 300, 500, 800, 2000, 5000)
  rvs_obs <- ci.fevd(fit_obs, alpha = 0.05, type = "return.level", return.period =␣
↪rperiods, method = "normal")
  colnames(rvs_obs) <- c("Obs_l", "Obs", "Obs_h") # Rename the col
  GEV_obs <- data.frame(cbind(rvs_obs, rperiods)) ## Make a datafram for ggplot

  ## Add the emipirical data
  rp_obs <- length(obs) / 1:length(obs) ## these are the (empirical) return periods␣
↪for the sorted datapoints
  obs_sorted <- sort(obs, decreasing = T) ## For example, the highest extreme has a␣
↪rp of 35 years, the second highest 17.5, third highest 11.7 etc.
  datapoints_obs <- data.frame(cbind(rp_obs, obs_sorted))


  rp_S5 <- length(ensemble) / 1:length(ensemble) # SEAS5 has return periods up to␣
↪3800 years
  ensemble_sorted <- sort(ensemble, decreasing = T)
  datapoints_S5 <- data.frame(cbind(rp_S5, ensemble_sorted))

  ## And plot
  cols <- c("UNSEEN" = "black", "OBS      " = "blue") ## for  the legend
  ggplot(data = datapoints_S5, aes(x = rp_S5)) +
    geom_point(aes(y = ensemble_sorted, col = "UNSEEN"), alpha = 0.5, size = 1) +
    geom_ribbon(data = GEV_obs, aes(ymin = Obs_l, ymax = Obs_h, x = rperiods, fill =
↪"OBS      "), alpha = 0.1) +
    geom_point(data = datapoints_obs, aes(x = rp_obs, y = obs_sorted, col = "OBS
↪"), size = 1) +
    scale_x_continuous(trans = "log10") +
    scale_fill_manual(name = "Data", values = cols) +
    scale_colour_manual(name = NULL, values = cols) +
    theme_classic() +
    theme(
      legend.position = c(.95, .05),
      legend.justification = c("right", "bottom"),
      legend.box.just = "right",
      legend.title = element_blank(),
      text = element_text(size = 11),
      axis.text = element_text(size = 11)
    ) +
    labs(title = main, y = y_lab, x = "Return period (years)") +
    if (is.finite(ylim)) {
      coord_cartesian(ylim = c(NA, ylim))
    }
}
```

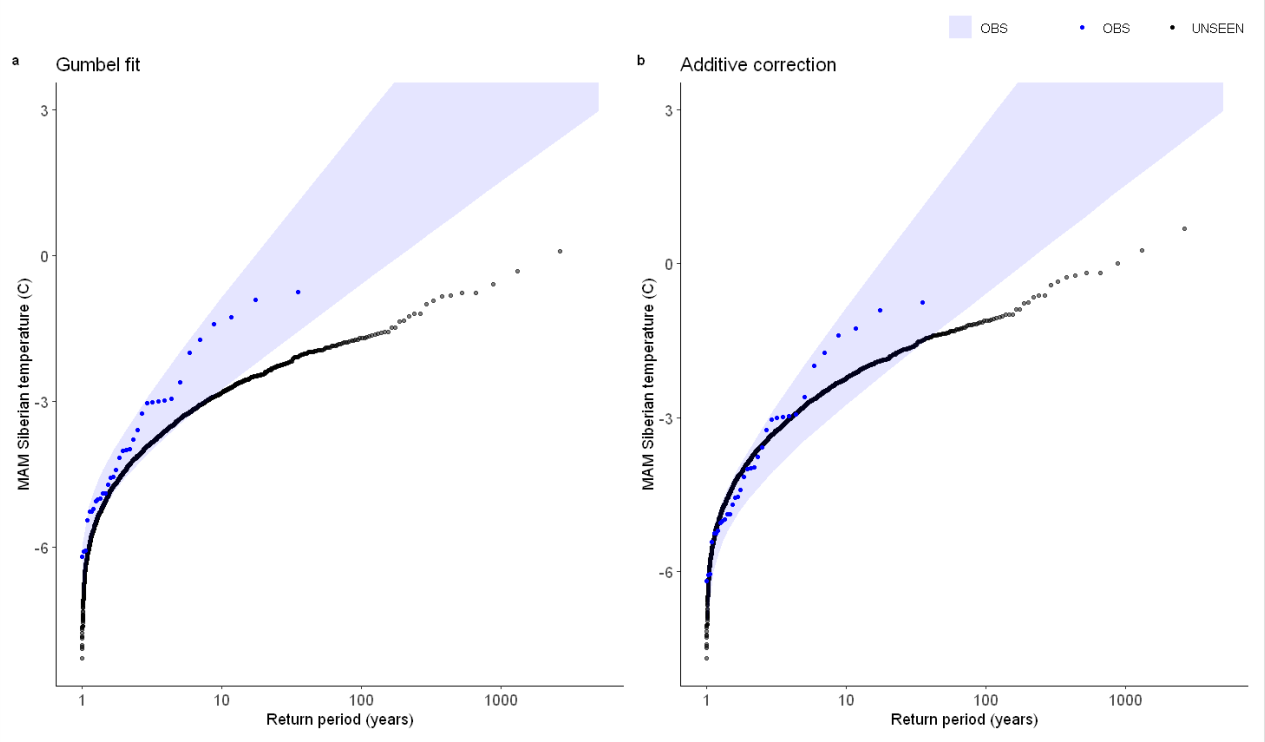
First, we fit a gumber and a GEV distribution (including shape parameter) to the observed extremes over Siberia. With a likelihood ratio test we show that the Gumbel distribution best describes the data.

[84]:
```
fit_obs_Gumbel <- fevd(x = ERA5_Siberia_events_zoomed_hindcast$t2m,
                       type = "Gumbel"
                      )
fit_obs_GEV <- fevd(x = ERA5_Siberia_events_zoomed_hindcast$t2m,
                       type = "GEV"
                      )
lr.test(fit_obs_Gumbel, fit_obs_GEV)
```

```
        Likelihood-ratio Test

data:  ERA5_Siberia_events_zoomed_hindcast$t2mERA5_Siberia_events_zoomed_hindcast$t2m
Likelihood-ratio = 0.21004, chi-square critical value = 3.8415, alpha =
0.0500, Degrees of Freedom = 1.0000, p-value = 0.6467
alternative hypothesis: greater
```

We show the gumbel plot for the observed (ERA5) and UNSEEN (SEAS5 hindcast data). This shows that the UNSEEN simulations are not within the uncertainty range of the observations. This has likely two reasons, illustrated in the evaluation section: there is some dependence between the events and there is too little variability within the UNSEEN ensemble.

[100]:
```
options(repr.plot.width = 12)
GEV_hindcast <- EVT_plot(ensemble = SEAS5_Siberia_events_zoomed_hindcast$t2m,
                         obs = ERA5_Siberia_events_zoomed_hindcast$t2m,
                         main = "Gumbel fit",
                         GEV_type = "Gumbel",
                         ylim = 3,
                         y_lab = 'MAM Siberian temperature (C)'
                        )
GEV_hindcast_corrected <- EVT_plot(ensemble = ensemble_biascor, #SEAS5_Siberia_events_
↪zoomed_hindcast$t2m,
                                   obs = ERA5_Siberia_events_zoomed_hindcast$t2m,
                                   main = "Additive correction",
                                   GEV_type = "Gumbel",
                                   ylim = 3,
                                   y_lab = 'MAM Siberian temperature (C)'
                                  )

ggarrange(GEV_hindcast, GEV_hindcast_corrected,
  labels = c("a", "b"), # ,"c","d"),
  common.legend = T,
  font.label = list(size = 10, color = "black", face = "bold", family = NULL),
  ncol = 2, nrow = 1
)
# GEV_hindcast
# GEV_hindcast_corrected
```

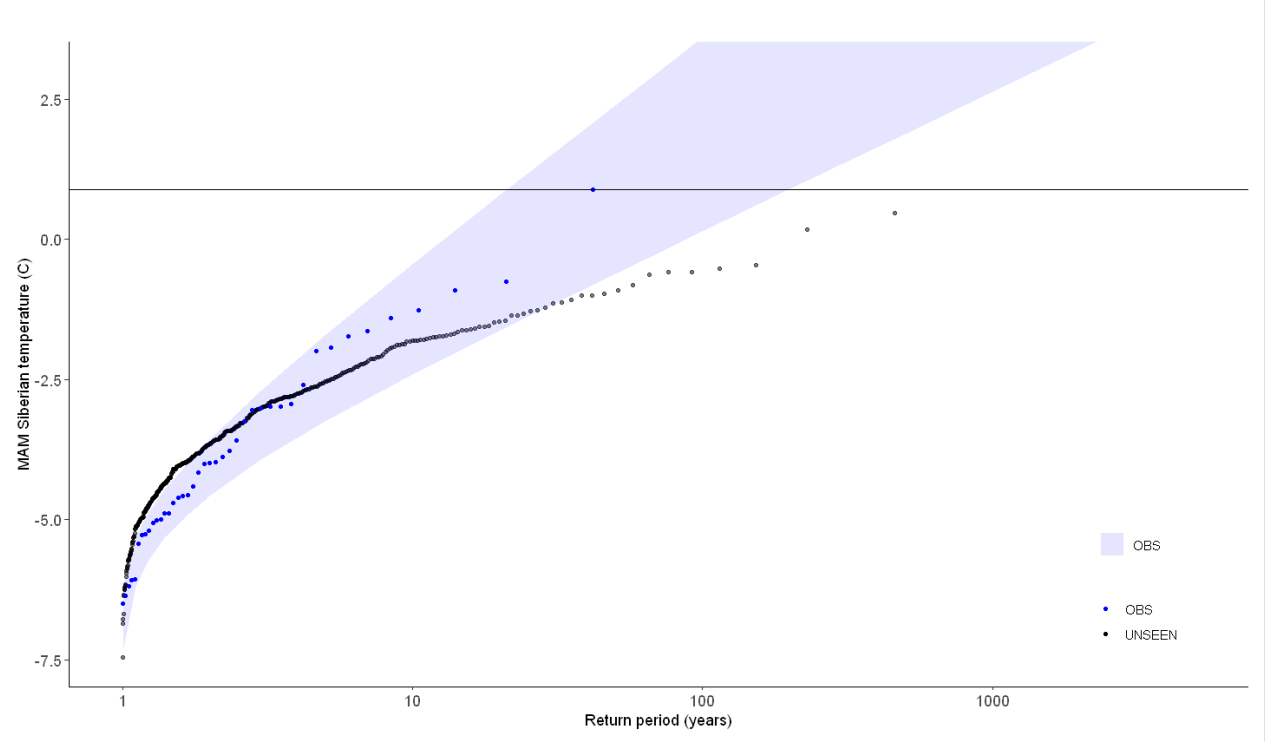


So what can we get out of it? What if we look at the operational forecast? Even if we cannot use the dataset as a whole to estimate the likelihood of occurrence, have events similar to the 2020 event occurred?

We select all archived SEAS5 (UNSEEN) events and all ERA5 (observed) events except for the 2020 event as reference.

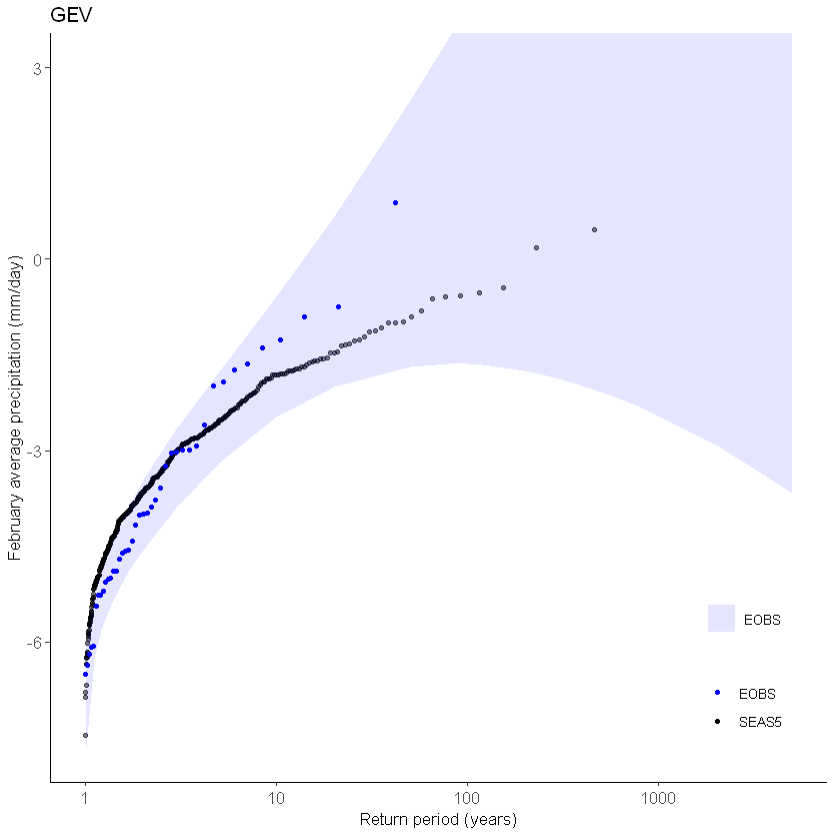
[146]:
```
ERA5_Siberia_events_zoomed_min1 <- ERA5_Siberia_events_zoomed[1:length(ERA5_Siberia_
↪events_zoomed$t2m)-1,]
ERA5_Siberia_events_zoomed_2020 <- ERA5_Siberia_events_zoomed[length(ERA5_Siberia_
↪events_zoomed$t2m),]
# ERA5_Siberia_events_zoomed_min1
# ERA5_Siberia_events_zoomed_2020
```

[151]:
```
GEV_forecasts <- EVT_plot(ensemble = SEAS5_Siberia_events_zoomed_forecasts$t2m,
                          obs = ERA5_Siberia_events_zoomed$t2m,
                          main = "",
                          GEV_type = "Gumbel",
                          ylim = 3,
                          y_lab = 'MAM Siberian temperature (C)'
                         ) # %>%
GEV_forecasts + geom_hline(yintercept = ERA5_Siberia_events_zoomed_2020$t2m)#,
#     color = "black", linetype = "dashed", size = 1
```

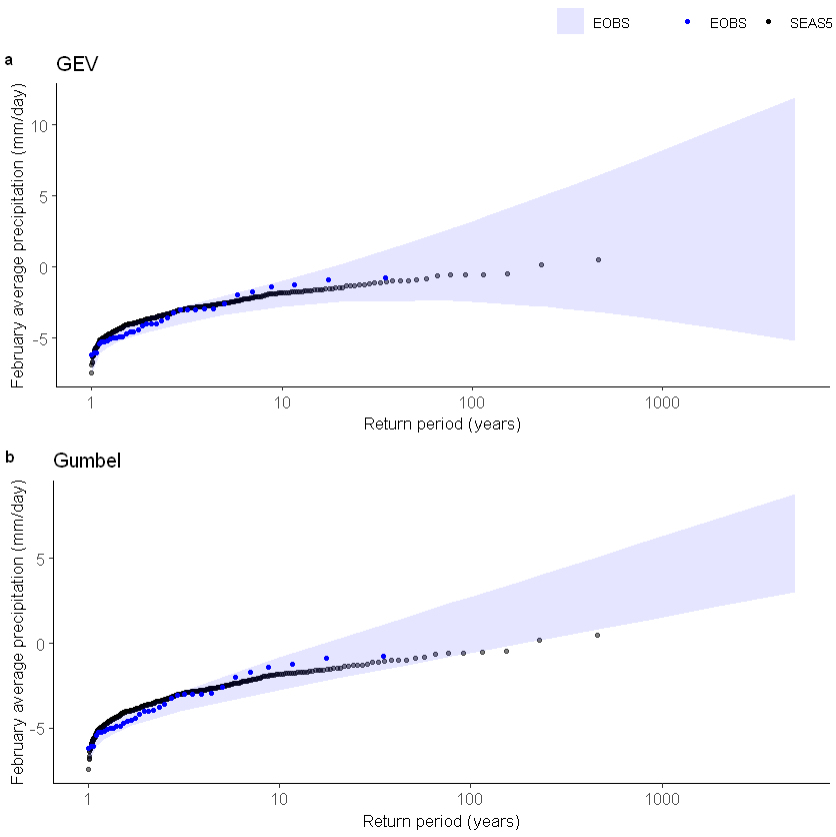


Plot the GEV distribution:

[63]:
```
GEV1 <- EVT_plot(ensemble = SEAS5_Siberia_events_zoomed_forecasts$t2m,
                 obs = ERA5_Siberia_events_zoomed$t2m,
                 main = "GEV",
                 GEV_type = "GEV",ylim = 3) # %>%
GEV1
```

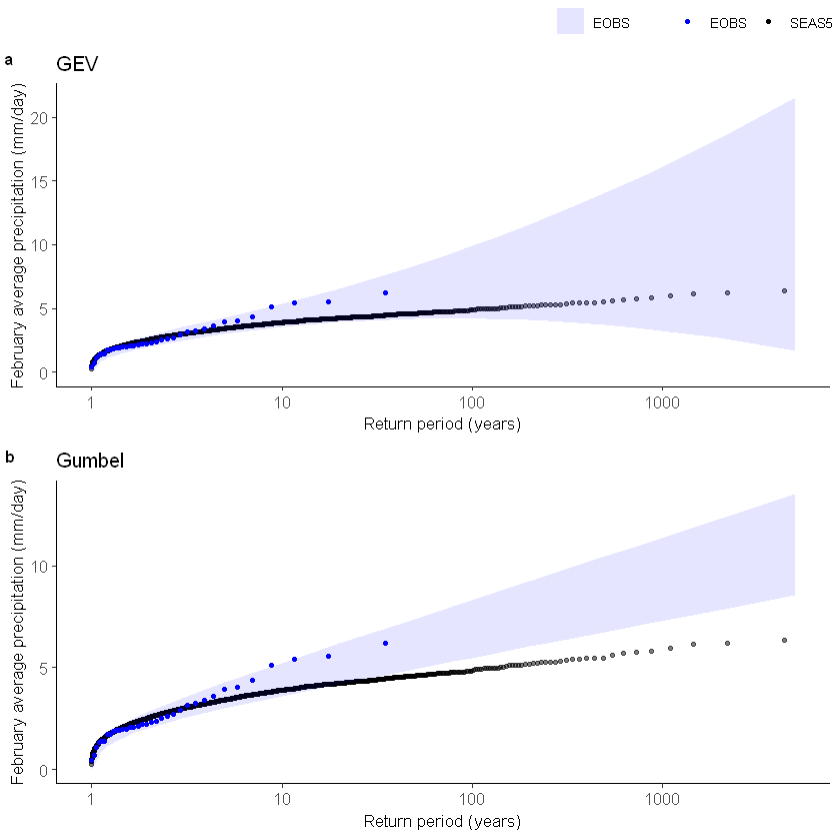


[57]:
```
Gumbel1 <- EVT_plot(ensemble = SEAS5_Siberia_events_zoomed_forecasts$t2m,
                    obs = ERA5_Siberia_events_zoomed_hindcast$t2m,
                    main = "Gumbel",
                    GEV_type = "Gumbel")

ggarrange(GEV1, Gumbel1,
  labels = c("a", "b"), # ,"c","d"),
  common.legend = T,
  font.label = list(size = 10, color = "black", face = "bold", family = NULL),
  ncol = 1, nrow = 2
) # %>%
# ggsave(filename = "graphs/Biascor.png",width =180,height = 180, units='mm',dpi=300)
```



And for the UK:

[56]:
```
GEV1 <- EVT_plot(ensemble = SEAS5_UK_weighted_df$tprate, obs = EOBS_UK_weighted_df_
↪hindcast$rr, main = "GEV", GEV_type = "GEV") # %>%
Gumbel1 <- EVT_plot(ensemble = SEAS5_UK_weighted_df$tprate, obs = EOBS_UK_weighted_df_
↪hindcast$rr, main = "Gumbel", GEV_type = "Gumbel") # %>%

ggarrange(GEV1, Gumbel1,
  labels = c("a", "b"), # ,"c","d"),
  common.legend = T,
  font.label = list(size = 10, color = "black", face = "bold", family = NULL),
  ncol = 1, nrow = 2
) # %>%
```



1.8 Global monthly temperature records in ERA5

Where have monthly average temperatures broken records across the world in 2020?

In this first section, we load required packages and modules

[29]:
```
##This is so variables get printed within jupyter
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

[2]:
```
##import packages
import os
import xarray as xr
import numpy as np
import matplotlib.pyplot as plt
import cartopy
import cartopy.crs as ccrs
import matplotlib.ticker as mticker

#for rank calculation
import bottleneck
```

[3]:
```
## this is to load our own function to retrieve ERA5,
## which is located in ../src/CDSretrieve.py
import sys
sys.path.append('../')
```

[4]:
```
##And here we load the module
import src.CDSretrieve as retrieve
```

[5]:
```
##We want the working directory to be the UNSEEN-open directory
pwd = os.getcwd() ##current working directory is UNSEEN-open/Notebooks/1.Download
pwd #print the present working directory
os.chdir(pwd+'/../') # Change the working directory to UNSEEN-open
os.getcwd() #print the working directory
```

[5]:
```
'/lustre/soge1/projects/ls/personal/timo/UNSEEN-open/Notebooks'
```

[5]:
```
'/lustre/soge1/projects/ls/personal/timo/UNSEEN-open'
```

1.8.1 Download ERA5

This section describes the retrieval of ERA5. We retrieve netcdf files of global monthly 2m temperature and 2m dewpoint temperature for each year over 1979-2020.

[39]:
```
retrieve.retrieve_ERA5(variables = ['2m_temperature','2m_dewpoint_temperature'],␣
↪folder = '../Siberia_example/')
;
```

[39]:
```
''
```

We load all files with xarray `open_mfdataset`. The latest 3 months in this dataset are made available through ERA5T, which might be slightly different to ERA5. In the downloaded file, an extra dimenions 'expver' indicates which data is ERA5 (`expver = 1`) and which is ERA5T (`expver = 5`). After retrieving and loading, I combine both ERA5 and ERA5T to create a dataset that runs until August 2020.

[10]:
```
ERA5 = xr.open_mfdataset('../Siberia_example/ERA5_????.nc',combine='by_coords') ##␣
↪open the data
ERA5#
```

[10]:
```
<xarray.Dataset>
Dimensions:    (expver: 2, latitude: 181, longitude: 360, time: 500)
Coordinates:
  * latitude   (latitude) float32 90.0 89.0 88.0 87.0 ... -88.0 -89.0 -90.0
  * longitude  (longitude) float32 -180.0 -179.0 -178.0 ... 177.0 178.0 179.0
  * expver     (expver) int32 1 5
  * time       (time) datetime64[ns] 1979-01-01 1979-02-01 ... 2020-08-01
Data variables:
    t2m        (time, latitude, longitude, expver) float32 dask.array<chunksize=(12,␣
↪181, 360, 2), meta=np.ndarray>
    d2m        (time, latitude, longitude, expver) float32 dask.array<chunksize=(12,␣
↪181, 360, 2), meta=np.ndarray>
Attributes:
    Conventions:  CF-1.6
    history:      2020-09-07 10:14:42 GMT by grib_to_netcdf-2.16.0: /opt/ecmw...
```

[14]:
```
ERA5_combine =ERA5.sel(expver=1).combine_first(ERA5.sel(expver=5))
ERA5_combine.load()
```

[14]:
```
<xarray.Dataset>
Dimensions:    (latitude: 181, longitude: 360, time: 500)
Coordinates:
  * latitude   (latitude) float32 90.0 89.0 88.0 87.0 ... -88.0 -89.0 -90.0
  * longitude  (longitude) float32 -180.0 -179.0 -178.0 ... 177.0 178.0 179.0
  * time       (time) datetime64[ns] 1979-01-01 1979-02-01 ... 2020-08-01
Data variables:
    t2m        (time, latitude, longitude) float32 244.7074 ... 214.79857
    d2m        (time, latitude, longitude) float32 241.76836 ... 211.0198
Attributes:
    Conventions:  CF-1.6
    history:      2020-09-07 10:14:42 GMT by grib_to_netcdf-2.16.0: /opt/ecmw...
```

1.8.2 Calculating the rank

We want to show for each month whether the recorded monthly average temperature for 2020 is the highest since 1979 (or second highest, etc.).

We first select only January months.

[15]:
```
ERA5_jan = ERA5_combine.sel(time=ERA5_combine['time.month'] == 1) ## Select only for␣
↪the i month
```

Then we calculate the rank of January average temperatures over the years. We rename the variable 't2m' into 'Temperature rank'.

[16]:
```
ERA5_jan_rank = ERA5_jan['t2m'].rank(dim = 'time')
ERA5_jan_rank = ERA5_jan_rank.rename('Temperature rank')
```

We now have calculated the rank in increasing order, i.e. the highest values has the highest rank. However, we want to show the highest rank being number 1, the second highest being number 2. Therefore, we invert the ranks and then we select the inverted rank of January 2020 average temperature within the January average temperatures of the other years. If January 2020 average temperature would be highest on record, the inverted rank will be 1. Second highest will be 2.

[17]:
```
ERA5_jan_rank_inverted = (len(ERA5_jan_rank.time) - ERA5_jan_rank + 1).sel(time='2020
↪')
ERA5_jan_rank_inverted
```

[17]:
```
<xarray.DataArray 'Temperature rank' (time: 1, latitude: 181, longitude: 360)>
array([[[25., 25., 25., ..., 25., 25., 25.],
        [23., 23., 23., ..., 23., 23., 23.],
        [25., 25., 25., ..., 25., 25., 25.],
        ...,
        [23., 23., 23., ..., 23., 23., 23.],
        [24., 24., 24., ..., 24., 24., 24.],
        [24., 24., 24., ..., 24., 24., 24.]]])
Coordinates:
  * latitude   (latitude) float32 90.0 89.0 88.0 87.0 ... -88.0 -89.0 -90.0
  * longitude  (longitude) float32 -180.0 -179.0 -178.0 ... 177.0 178.0 179.0
  * time       (time) datetime64[ns] 2020-01-01
```

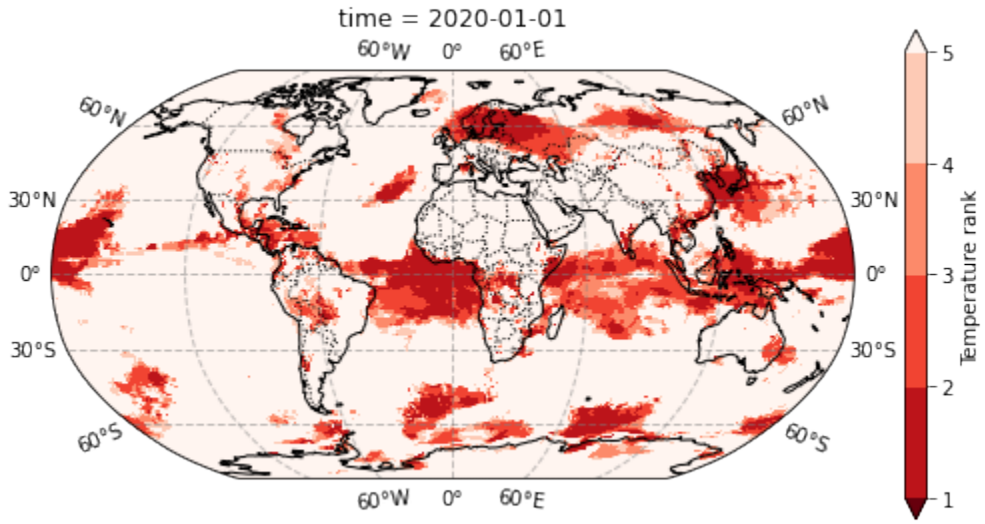
1.8.3 Plotting

We define a function to plot the data on a global map:

[18]:
```
def Global_plot(ERA5_i_rank_inverted):
    fig, ax = plt.subplots(figsize=(9, 4.5))
    ax = plt.axes(projection=ccrs.Robinson())
    ERA5_i_rank_inverted.plot(
        ax=ax,
        transform=ccrs.PlateCarree(),
        levels=[1, 2, 3, 4, 5],
        extend='both',
        colors=plt.cm.Reds_r)

    ax.add_feature(cartopy.feature.BORDERS, linestyle=':')
    ax.coastlines(
        resolution='110m')  #Currently can be one of "110m", "50m", and "10m".
    gl = ax.gridlines(crs=ccrs.PlateCarree(),
                      draw_labels=True,
                      linewidth=1,
                      color='gray',
                      alpha=0.5,
                      linestyle='--')
#     gl.top_labels = False
#     gl.right_labels = False
```

And plot!

[19]:
```
Global_plot(ERA5_jan_rank_inverted)
```



And zoom in for Siberia. We define a new plot:

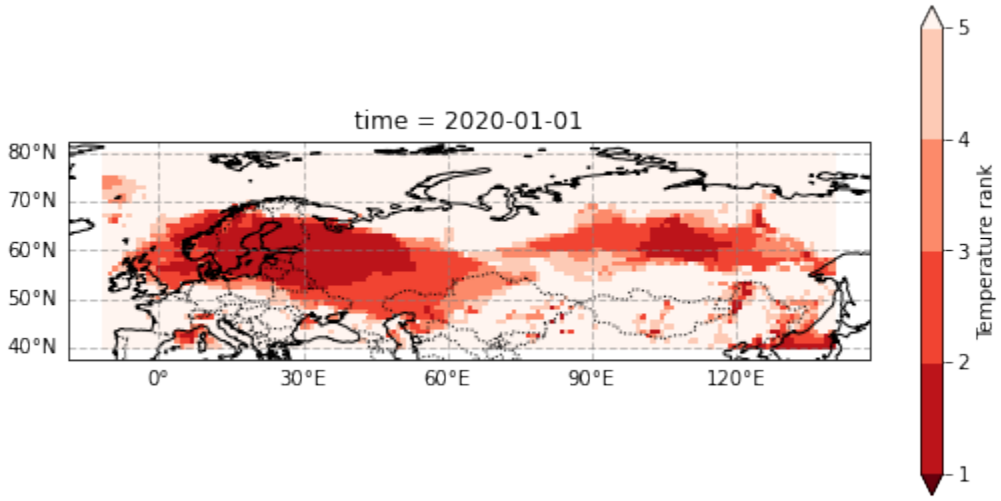
[55]:
```
def Siberia_plot(ERA5_i_rank_inverted):
    fig, ax = plt.subplots(figsize=(9, 4.5))
    ax = plt.axes(projection=ccrs.PlateCarree(central_longitude=50.0))
    ERA5_i_rank_inverted.plot(
        ax=ax,
        transform=ccrs.PlateCarree(),
        levels=[1, 2, 3, 4, 5],
```

(continues on next page)

(continued from previous page)

```
        extend='both',
        colors=plt.cm.Reds_r)

    ax.add_feature(cartopy.feature.BORDERS, linestyle=':')
    ax.coastlines(resolution='50m')
    gl = ax.gridlines(crs=ccrs.PlateCarree(),
                      draw_labels=True,
                      linewidth=1,
                      color='gray',
                      alpha=0.5,
                      linestyle='--')
    gl.top_labels = False
    gl.right_labels = False
```

[56]:
```
Siberia_plot(ERA5_jan_rank_inverted.sel(longitude = slice(-11,140), latitude =␣
↪slice(80,40)))
```



1.8.4 Loop over Jan-Aug

1.8.5 Create the gif

We use ImageMagick and run it from the command line. See this CMS notebook for more info on creating gifs.

[58]:
```
!convert -delay 60 ../Siberia_example/plots/Global*png graphs/Global_Animation_01.gif
```

[60]:
```
!convert -delay 60 ../Siberia_example/plots/Siberia*png graphs/Siberia_Animation_01.
↪gif
```

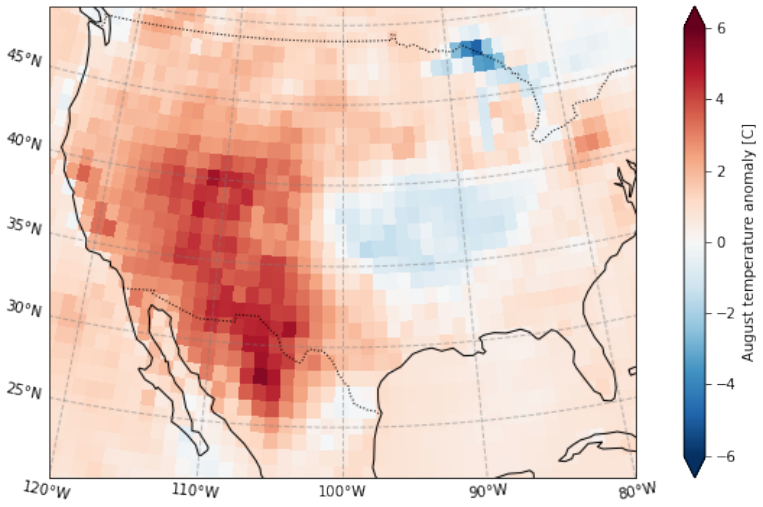
And show the gif in jupyter notebook with

```
![Global Temperature records 2020](../graphs/Global_Animation_01.gif
"Records2020")
```

Same for the Siberian temperature records: `![Siberian Temperature records 2020](../graphs/Siberia_Animation_01.gif "Records2020")`

1.9 California august temperature anomaly

How anomalous was the August 2020 average temperature?



In this first section, we load required packages and modules

[1]:
```
##This is so variables get printed within jupyter
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

[2]:
```
##import packages
import os
import xarray as xr
import numpy as np
import matplotlib.pyplot as plt
import cartopy
import cartopy.crs as ccrs
import matplotlib.ticker as mticker

#for rank calculation
# import bottleneck
```

[3]:
```
os.chdir(os.path.abspath('../../'))
```

1.9.1 Load ERA5

We have retrieve netcdf files of global monthly 2m temperature and 2m dewpoint temperature for each year over 1979-2020.

We load all files with xarray `open_mfdataset`.

[4]:
```
ERA5 = xr.open_mfdataset('E:/PhD/California_example/ERA5/ERA5_????.nc',combine='by_
↪coords') ## open the data
ERA5#
```

[4]:
```
<xarray.Dataset>
Dimensions:    (latitude: 51, longitude: 61, time: 42)
Coordinates:
  * longitude  (longitude) float32 -130.0 -129.0 -128.0 ... -72.0 -71.0 -70.0
  * latitude   (latitude) float32 70.0 69.0 68.0 67.0 ... 23.0 22.0 21.0 20.0
  * time       (time) datetime64[ns] 1979-08-01 1980-08-01 ... 2020-08-01
Data variables:
    t2m        (time, latitude, longitude) float32 dask.array<chunksize=(1, 51, 61),␣
↪meta=np.ndarray>
    d2m        (time, latitude, longitude) float32 dask.array<chunksize=(1, 51, 61),␣
↪meta=np.ndarray>
Attributes:
    Conventions:  CF-1.6
    history:      2020-10-01 23:23:34 GMT by grib_to_netcdf-2.16.0: /opt/ecmw...
```

1.9.2 Calculating the anomaly

We want to show how anomalous the recorded monthly average temperature for 2020 is compared to the 1979-2010 average.

[13]:
```
ERA5_anomaly = ERA5['t2m'] - ERA5['t2m'].sel(time=slice('1979','2010')).mean('time')
ERA5_anomaly.attrs = {
    'long_name': 'August temperature anomaly',
    'units': 'C'
}
ERA5_sd_anomaly = ERA5_anomaly / ERA5['t2m'].std('time')
ERA5_sd_anomaly.attrs = {
    'long_name': 'August temperature standardized anomaly',
    'units': '-'
}
```

1.9.3 Plotting

We define a function to plot the data on a global map:

[14]:
```
def plot_California(ERA5_input):

    extent = [-120, -80, 20, 50]
    central_lon = np.mean(extent[:2])
    central_lat = np.mean(extent[2:])

    plt.figure(figsize=(12, 6))
    ax = plt.axes(projection=ccrs.AlbersEqualArea(central_lon, central_lat))
    ax.set_extent(extent)
```

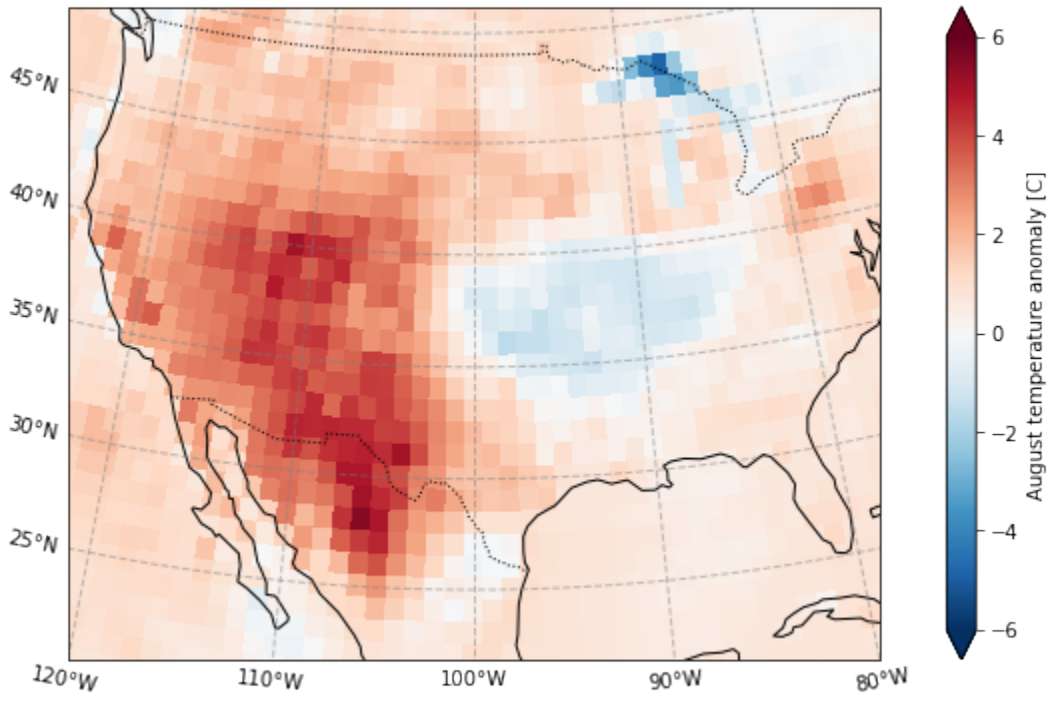
(continues on next page)

(continued from previous page)

```
    ERA5_input.plot(
        ax=ax,
        transform=ccrs.PlateCarree(),
    #     levels=[1, 2, 3, 4, 5],
        extend='both')#,
    #     colors=plt.cm.Reds_r)

    ax.add_feature(cartopy.feature.BORDERS, linestyle=':')
    ax.coastlines(
        resolution='110m')  #Currently can be one of "110m", "50m", and "10m".
    ax.set_title('')
    gl = ax.gridlines(crs=ccrs.PlateCarree(),
                      draw_labels=True,
                      linewidth=1,
                      color='gray',
                      alpha=0.5,
                      linestyle='--')
    gl.top_labels = False
    gl.right_labels = False
```

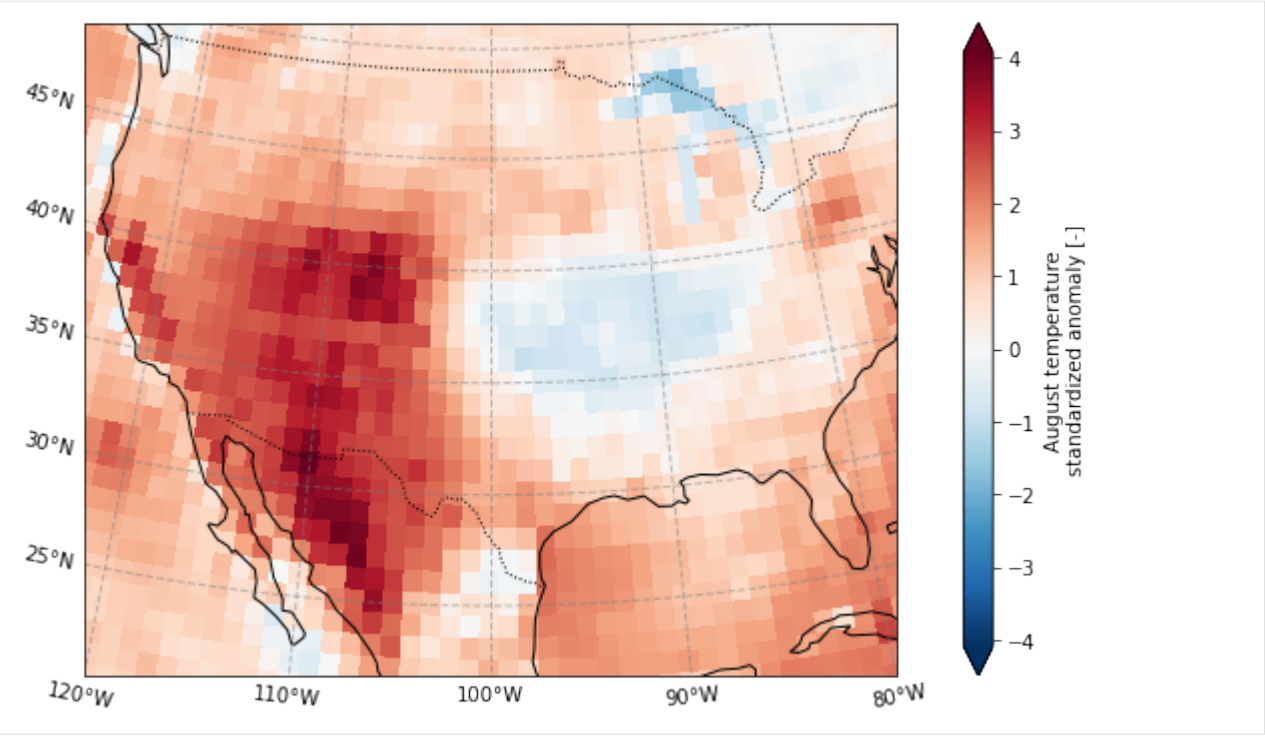
Temperature anomaly

[15]:
```
plot_California(ERA5_anomaly.sel(time = '2020'))
plt.savefig('graphs/California_anomaly.png')
```



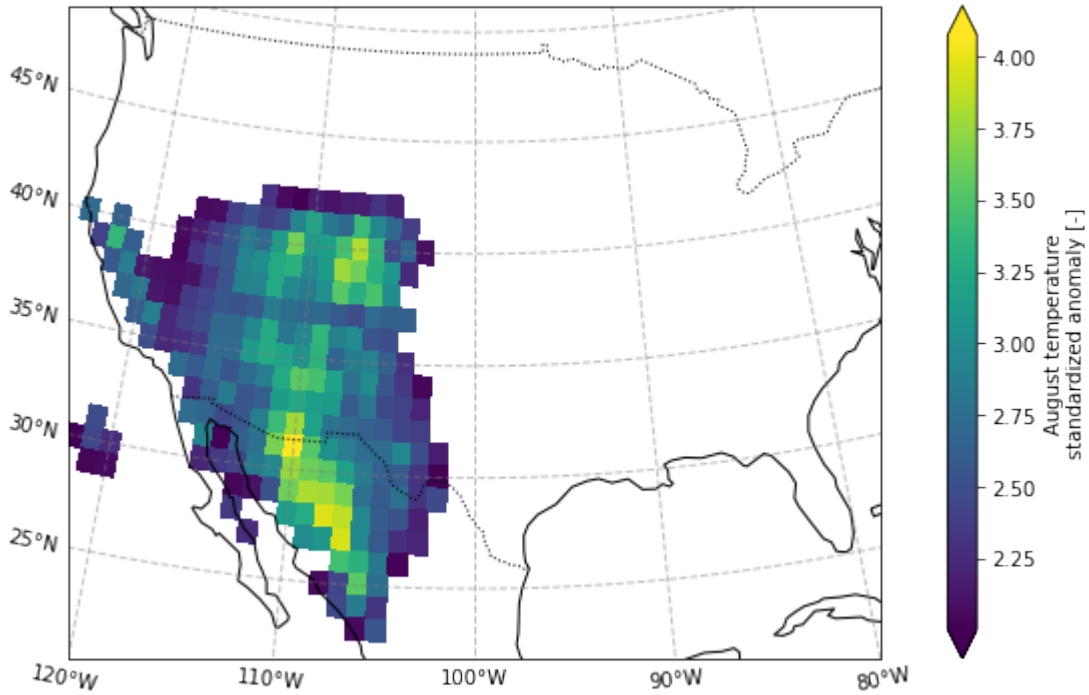
Plot the standardized anomaly

[16]:
```
plot_California(ERA5_sd_anomaly.sel(time = '2020'))
plt.savefig('graphs/California_sd_anomaly.png')
```



Define mask as higher than 2 standard deviation anomaly

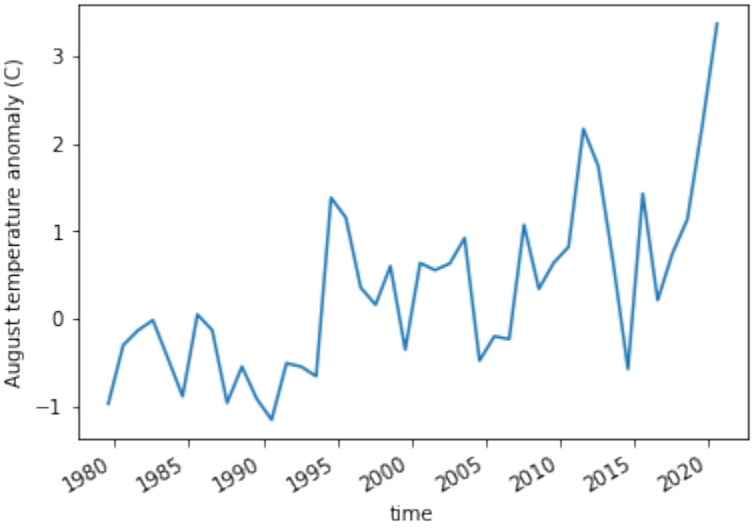
[17]:
```
ERA5_masked = (ERA5_sd_anomaly.
               sel(longitude = slice(-125,-100),
                  latitude = slice(45,20)).
               where(ERA5_sd_anomaly.sel(time = '2020').
                     squeeze('time')>2)
              )

# ERA5_masked
plot_California(ERA5_masked.sel(time = '2020'))
```



[21]:
```
ERA5_anomaly_timeseries = ERA5_anomaly.sel(longitude = slice(-119,-100)).where(ERA5_
↪sd_anomaly.sel(time = '2020').squeeze('time')>2).mean(['longitude','latitude'])
ERA5_anomaly_timeseries.plot()
plt.ylabel('August temperature anomaly (C)')
plt.savefig('graphs/California_anomaly_timeseries.png')
```

[21]:
```
[<matplotlib.lines.Line2D at 0x241db54c790>]
```

[21]:
```
Text(0, 0.5, 'August temperature anomaly (C)')
```



1.10 February and April 2020 precipitation anomalies

In this notebook, we will analyze precipitation anomalies of February and April 2020, which seemed to be very contrasting in weather. We use the EOBS dataset.

1.10.1 Import packages

[1]:
```
##This is so variables get printed within jupyter
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

[2]:
```
##import packages
import os
import xarray as xr
import numpy as np
import matplotlib.pyplot as plt
import cartopy
import cartopy.crs as ccrs
import matplotlib.ticker as mticker
```

[3]:
```
##We want the working directory to be the UNSEEN-open directory
pwd = os.getcwd() ##current working directory is UNSEEN-open/Notebooks/1.Download
pwd #print the present working directory
os.chdir(pwd+'/../') # Change the working directory to UNSEEN-open
os.getcwd() #print the working directory
```

[3]:
```
'/lustre/soge1/projects/ls/personal/timo/UNSEEN-open/Notebooks/1.Download'
```

[3]:
```
'/lustre/soge1/projects/ls/personal/timo/UNSEEN-open'
```

1.10.2 Load EOBS

I downloaded EOBS (from 1950 - 2019) and the most recent EOBS data (2020) here. Note, you have to register as E-OBS user.

The data has a daily timestep. I resample the data into monthly average mm/day. I chose not to use the total monthly precipitation because of leap days.

[4]:
```
EOBS = xr.open_dataset('../UK_example/EOBS/rr_ens_mean_0.25deg_reg_v20.0e.nc') ##␣
↪open the data
EOBS = EOBS.resample(time='1m').mean() ## Monthly averages
# EOBS = EOBS.sel(time=EOBS['time.month'] == 2) ## Select only February
EOBS
```

```
/soge-home/users/cenv0732/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/xarray/
↪core/nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

[4]:
```
<xarray.Dataset>
Dimensions:    (latitude: 201, longitude: 464, time: 835)
Coordinates:
  * time       (time) datetime64[ns] 1950-01-31 1950-02-28 ... 2019-07-31
  * latitude   (latitude) float64 25.38 25.62 25.88 26.12 ... 74.88 75.12 75.38
  * longitude  (longitude) float64 -40.38 -40.12 -39.88 ... 74.88 75.12 75.38
```

(continues on next page)

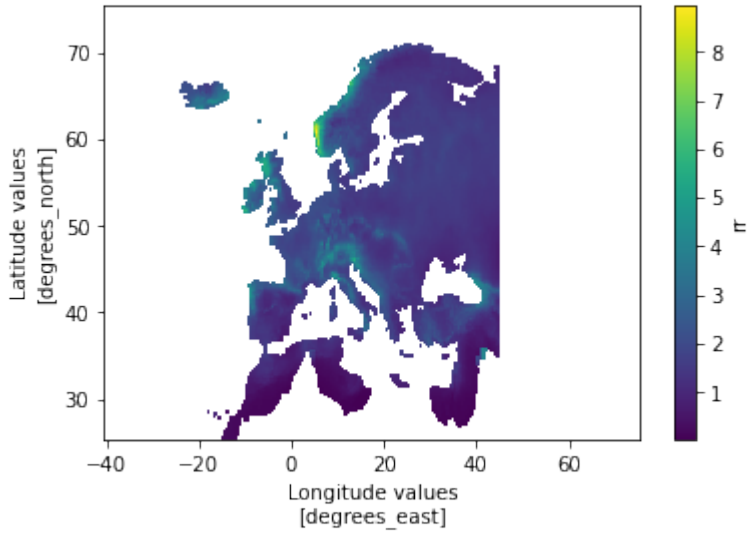
(continued from previous page)

```
Data variables:
    rr          (time, latitude, longitude) float32 nan nan nan ... nan nan nan
```

Here I define the attributes, that xarray uses when plotting

[5]:
```
EOBS['rr'].attrs = {'long_name': 'rainfall',  ##Define the name
 'units': 'mm/day', ## unit
 'standard_name': 'thickness_of_rainfall_amount'} ## original name, not used
EOBS['rr'].mean('time').plot() ## and show the 1950-2019 average February␣
↪precipitation
```

[5]:
```
<matplotlib.collections.QuadMesh at 0x7f002a87ad90>
```



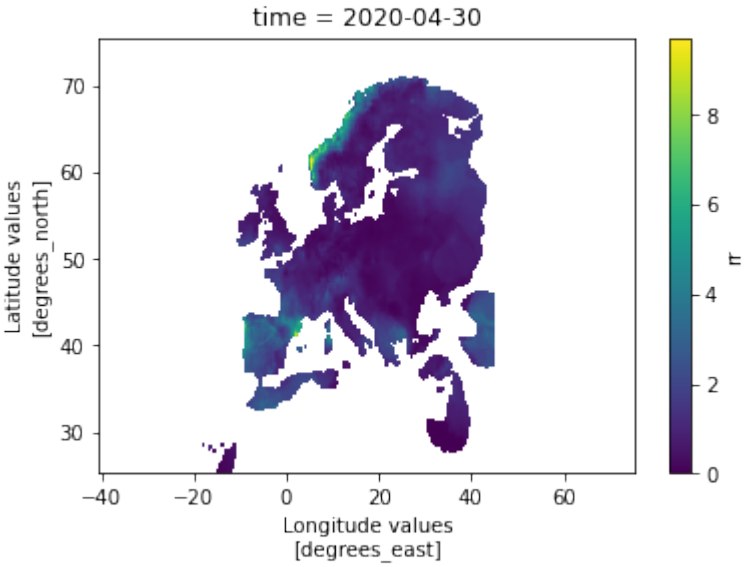
The 2020 data file is separate and needs the same preprocessing:

[6]:
```
EOBS2020 = xr.open_dataset('../UK_example/EOBS/rr_0.25deg_day_2020_grid_ensmean.nc.1
↪') #open
EOBS2020 = EOBS2020.resample(time='1m').mean() #Monthly mean
EOBS2020['rr'].sel(time='2020-04').plot() #show map
EOBS2020 ## display dataset
```

```
/soge-home/users/cenv0732/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/xarray/
↪core/nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

[6]:
```
<matplotlib.collections.QuadMesh at 0x7f002a76d8e0>
```

[6]:
```
<xarray.Dataset>
Dimensions:    (latitude: 201, longitude: 464, time: 12)
Coordinates:
  * time       (time) datetime64[ns] 2020-01-31 2020-02-29 ... 2020-12-31
  * latitude   (latitude) float64 25.38 25.62 25.88 26.12 ... 74.88 75.12 75.38
  * longitude  (longitude) float64 -40.38 -40.12 -39.88 ... 74.88 75.12 75.38
Data variables:
    rr         (time, latitude, longitude) float32 nan nan nan ... nan nan nan
```



1.10.3 Plot the 2020 event

I calculate the anomaly (deviation from the mean in mm/d) and divide this by the standard deviation to obtain the standardized anomalies.

[46]:
```
EOBS2020_anomaly = EOBS2020['rr'].groupby('time.month') - EOBS['rr'].groupby('time.
↪month').mean('time')
EOBS2020_anomaly

EOBS2020_sd_anomaly = EOBS2020_anomaly.groupby('time.month') / EOBS['rr'].groupby(
↪'time.month').std('time')

EOBS2020_sd_anomaly.attrs = {
    'long_name': 'Monthly precipitation standardized anomaly',
    'units': '-'
}

EOBS2020_sd_anomaly
```

```
/soge-home/users/cenv0732/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/xarray/
↪core/nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

[46]:
```
<xarray.DataArray 'rr' (time: 12, latitude: 201, longitude: 464)>
array([[[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
```

(continues on next page)

(continued from previous page)

```
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       ...,

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]]], dtype=float32)
Coordinates:
  * latitude   (latitude) float64 25.38 25.62 25.88 26.12 ... 74.88 75.12 75.38
  * longitude  (longitude) float64 -40.38 -40.12 -39.88 ... 74.88 75.12 75.38
  * time       (time) datetime64[ns] 2020-01-31 2020-02-29 ... 2020-12-31
    month      (time) int64 1 2 3 4 5 6 7 8 9 10 11 12
```

```
/soge-home/users/cenv0732/.conda/envs/UNSEEN-open/lib/python3.8/site-packages/numpy/
↪lib/nanfunctions.py:1666: RuntimeWarning: Degrees of freedom <= 0 for slice.
  var = nanvar(a, axis=axis, dtype=dtype, out=out, ddof=ddof,
```

[46]:
```
<xarray.DataArray 'rr' (time: 12, latitude: 201, longitude: 464)>
array([[[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
```

(continues on next page)

(continued from previous page)

```
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       ...,

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]]], dtype=float32)
Coordinates:
  * latitude   (latitude) float64 25.38 25.62 25.88 26.12 ... 74.88 75.12 75.38
  * longitude  (longitude) float64 -40.38 -40.12 -39.88 ... 74.88 75.12 75.38
  * time       (time) datetime64[ns] 2020-01-31 2020-02-29 ... 2020-12-31
    month      (time) int64 1 2 3 4 5 6 7 8 9 10 11 12
Attributes:
    long_name:  Monthly precipitation standardized anomaly
    units:      -
```

I select February and April (tips on how to select this are appreciated)

[38]:
```
EOBS2020_sd_anomaly
# EOBS2020_sd_anomaly.sel(time = ['2020-02','2020-04']) ## Dont know how to select␣
↪this by label?
EOBS2020_sd_anomaly[[1,3],:,:] ## Dont know how to select this by label?
```

[38]:
```
<xarray.DataArray 'rr' (time: 12, latitude: 201, longitude: 464)>
```

(continues on next page)

(continued from previous page)

```
array([[[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       ...,

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]]], dtype=float32)
Coordinates:
  * latitude   (latitude) float64 25.38 25.62 25.88 26.12 ... 74.88 75.12 75.38
  * longitude  (longitude) float64 -40.38 -40.12 -39.88 ... 74.88 75.12 75.38
  * time       (time) datetime64[ns] 2020-01-31 2020-02-29 ... 2020-12-31
    month      (time) int64 1 2 3 4 5 6 7 8 9 10 11 12
Attributes:
    long_name:  Monthly precipitation standardized anomaly
    units:      -
```

[38]:
```
<xarray.DataArray 'rr' (time: 2, latitude: 201, longitude: 464)>
array([[[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]],

       [[nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        ...,
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan],
        [nan, nan, nan, ..., nan, nan, nan]]], dtype=float32)
Coordinates:
  * latitude   (latitude) float64 25.38 25.62 25.88 26.12 ... 74.88 75.12 75.38
  * longitude  (longitude) float64 -40.38 -40.12 -39.88 ... 74.88 75.12 75.38
  * time       (time) datetime64[ns] 2020-02-29 2020-04-30
    month      (time) int64 2 4
Attributes:
    long_name:  Monthly precipitation standardized anomaly
    units:      -
```

And plot using cartopy!

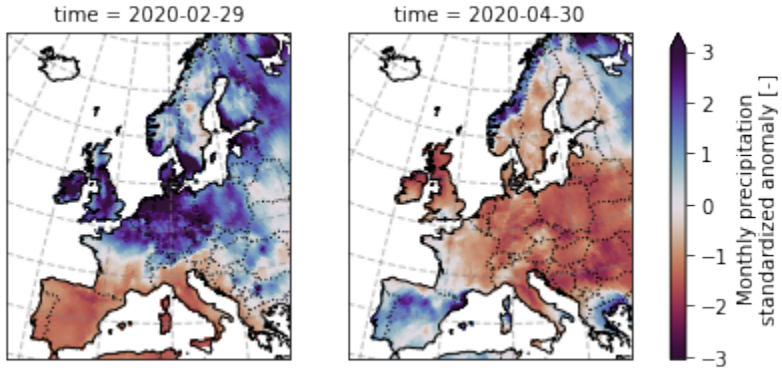
[58]:
```
EOBS_plots = EOBS2020_sd_anomaly[[1, 3], :, :].plot(
    transform=ccrs.PlateCarree(),
    robust=True,
    col='time',
    cmap=plt.cm.twilight_shifted_r,
    subplot_kws={'projection': ccrs.EuroPP()})

for ax in EOBS_plots.axes.flat:
    ax.add_feature(cartopy.feature.BORDERS, linestyle=':')
    ax.coastlines(resolution='50m')
    gl = ax.gridlines(crs=ccrs.PlateCarree(),
                      draw_labels=False,
                      linewidth=1,
                      color='gray',
                      alpha=0.5,
                      linestyle='--')

# plt.savefig('graphs/February_April_2020_precipAnomaly.png', dpi=300)
```

[58]:
```
<cartopy.mpl.feature_artist.FeatureArtist at 0x7efffc1af760>
```

[58]:
```
<cartopy.mpl.feature_artist.FeatureArtist at 0x7effcc6d4310>
```

[58]:
```
<cartopy.mpl.feature_artist.FeatureArtist at 0x7effb064a640>
```

[58]:
```
<cartopy.mpl.feature_artist.FeatureArtist at 0x7effb0603310>
```



1.11 Using EOBS + upscaling

Here we explore how to best extract areal averaged precipitation and test this for UK precipitation within SEAS5 and EOBS. The code is inspired on Matteo De Felice's blog – credits to him!

We create a mask for all 241 countries within Regionmask, that has predefined countries from Natural Earth datasets (shapefiles). We use the mask to go from gridded precipitation to country-averaged timeseries. We regrid EOBS to the SEAS5 grid so we can select the same grid cells in calculating the UK average for both datasets. The country outline would not be perfect, but the masks would be the same so the comparison would be fair.

I use the xesmf package for upscaling, a good example can be found in this notebook.

1.11.1 Import packages

We need the packages regionmask for masking and xesmf for regridding. I cannot install xesmf into the UNSEEN-open environment without breaking my environment, so in this notebook I use a separate 'upscale' environment, as suggested by this issue. I use the packages esmpy=7.1.0 xesmf=0.2.1 regionmask cartopy matplotlib xarray numpy netcdf4.

[1]:
```
##This is so variables get printed within jupyter
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

[2]:
```
##import packages
import os
import xarray as xr
import numpy as np
import matplotlib.pyplot as plt
import cartopy
import cartopy.crs as ccrs
import matplotlib.ticker as mticker

import regionmask        # Masking
import xesmf as xe       # Regridding
```

[3]:
```
##We want the working directory to be the UNSEEN-open directory
pwd = os.getcwd() ##current working directory is UNSEEN-open/Notebooks/1.Download
pwd #print the present working directory
```

(continues on next page)

(continued from previous page)

```
os.chdir(pwd+'/../../') # Change the working directory to UNSEEN-open
os.getcwd() #print the working directory
```

[3]:
```
'/lustre/soge1/projects/ls/personal/timo/UNSEEN-open/Notebooks/1.Download'
```

[3]:
```
'/lustre/soge1/projects/ls/personal/timo/UNSEEN-open'
```

1.11.2 Load SEAS5 and EOBS

From CDS, we retrieve SEAS5 and merge the retrieved files in preprocessing. We create a netcdf file containing the dimensions lat, lon, time (35 years), number (25 ensembles) and leadtime (5 initialization months).

[4]:
```
SEAS5 = xr.open_dataset('../UK_example/SEAS5/SEAS5.nc')
SEAS5
```

[4]:
```
<xarray.Dataset>
Dimensions:     (latitude: 11, leadtime: 5, longitude: 14, number: 25, time: 35)
Coordinates:
  * longitude   (longitude) float32 -11.0 -10.0 -9.0 -8.0 ... -1.0 0.0 1.0 2.0
  * time        (time) datetime64[ns] 1982-02-01 1983-02-01 ... 2016-02-01
  * latitude    (latitude) float32 60.0 59.0 58.0 57.0 ... 53.0 52.0 51.0 50.0
  * number      (number) int32 0 1 2 3 4 5 6 7 8 9 ... 16 17 18 19 20 21 22 23 24
  * leadtime    (leadtime) int64 2 3 4 5 6
Data variables:
    tprate      (leadtime, time, number, latitude, longitude) float32 ...
Attributes:
    Conventions:  CF-1.6
    history:      2020-05-13 14:49:43 GMT by grib_to_netcdf-2.16.0: /opt/ecmw...
```

And load EOBS netcdf with only February precipitation, resulting in 71 values, one for each year within 1950 - 2020 over the European domain (25N-75N x 40W-75E).

[5]:
```
EOBS = xr.open_dataset('../UK_example/EOBS/EOBS.nc')
EOBS
```

[5]:
```
<xarray.Dataset>
Dimensions:     (latitude: 201, longitude: 464, time: 71)
Coordinates:
  * longitude   (longitude) float64 -40.38 -40.12 -39.88 ... 74.88 75.12 75.38
  * latitude    (latitude) float64 25.38 25.62 25.88 26.12 ... 74.88 75.12 75.38
  * time        (time) datetime64[ns] 1950-02-28 1951-02-28 ... 2020-02-29
Data variables:
    rr          (time, latitude, longitude) float32 ...
```

1.11.3 Masking

Here we load the countries and create a mask for SEAS5 and for EOBS.

Regionmask has predefined countries from Natural Earth datasets (shapefiles).

[7]:
```
countries = regionmask.defined_regions.natural_earth.countries_50
countries
```

[7]:
```
241 'Natural Earth Countries: 50m' Regions (http://www.naturalearthdata.com)
ZW ZM YE VN VE V VU UZ UY FSM MH MP VI GU AS PR US GS IO SH PN AI FK KY BM VG TC MS␣
↪JE GG IM GB AE UA UG TM TR TN TT TO TG TL TH TZ TJ TW SYR CH S SW SR SS SD LK E KR␣
↪ZA SO SL SB SK SLO SG SL SC RS SN SA ST RSM WS VC LC KN RW RUS RO QA P PL PH PF PY␣
↪PG PA PW PK OM N KP NG NE NI NZ NU CK NL AW CW NP NR NA MZ MA WS ME MN MD MC MX MU␣
↪MR M ML MV MY MW MG MK L LT FL LY LR LS LB LV LA KG KW KO KI KE KZ J J J I IS PAL␣
↪IRL IRQ IR ID HND TT HU HN HT GY GW GN GT GD GR GH D GE GM GA F PM WF MF BL PF␣
↪NC TF AI FIN FJ ET EST ER GQ SV EG EC DO DM DJ GL FO DK CZ CN CY CU HR CI CR DRC CG␣
↪KM CO CN MO HK CL TD CF CV CA CM KH MM BI BF BG BN BR BW BiH BO BT BJ BZ B BY BB BD␣
↪BH BS AZ A AU IOT HM NF AU ARM AR AG AO AND DZ AL AF SG AQ SX
```

(continues on next page)

(continued from previous page)

Now we create the mask for the SEAS5 grid. Only one timestep is needed to create the mask. This mask will lateron be used to mask all the timesteps.

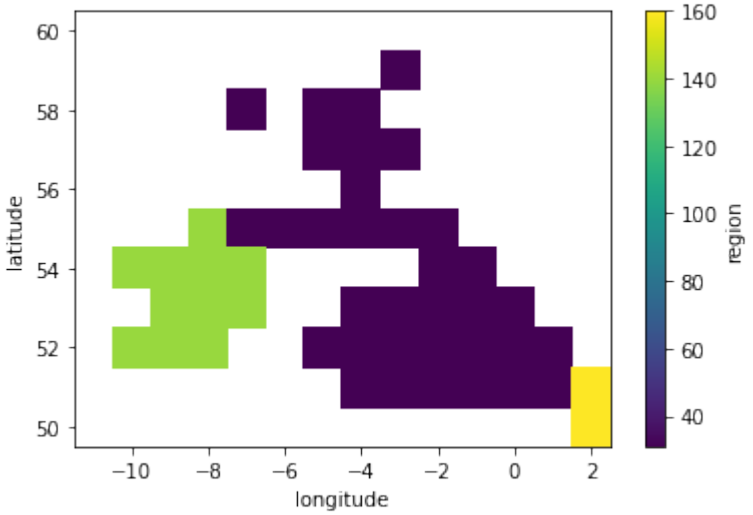
[8]:
```
SEAS5_mask = countries.mask(SEAS5.sel(leadtime=2, number=0, time='1982'),
                            lon_name='longitude',
                            lat_name='latitude')
```

And create a plot to illustrate what the mask looks like. The mask just indicates for each gridcell what country the gridcell belongs to.

[9]:
```
SEAS5_mask
SEAS5_mask.plot()
```

[9]:
```
<xarray.DataArray 'region' (latitude: 11, longitude: 14)>
array([[ nan,  nan,  nan,  nan,  nan,  nan,  nan,  nan,  nan,  nan,  nan,
         nan,  nan,  nan],
       [ nan,  nan,  nan,  nan,  nan,  nan,  nan,  nan,  31.,  nan,  nan,
         nan,  nan,  nan],
       [ nan,  nan,  nan,  nan,  31.,  nan,  31.,  31.,  nan,  nan,  nan,
         nan,  nan,  nan],
       [ nan,  nan,  nan,  nan,  nan,  nan,  31.,  31.,  31.,  nan,  nan,
         nan,  nan,  nan],
       [ nan,  nan,  nan,  nan,  nan,  nan,  nan,  31.,  nan,  nan,  nan,
         nan,  nan,  nan],
       [ nan,  nan,  nan, 140.,  31.,  31.,  31.,  31.,  31.,  31.,  nan,
         nan,  nan,  nan],
       [ nan, 140., 140., 140., 140.,  nan,  nan,  nan,  nan,  31.,  31.,
         nan,  nan,  nan],
       [ nan,  nan, 140., 140., 140.,  nan,  nan,  31.,  31.,  31.,  31.,
         31.,  nan,  nan],
       [ nan, 140., 140., 140.,  nan,  nan,  31.,  31.,  31.,  31.,  31.,
         31.,  31.,  nan],
       [ nan,  nan,  nan,  nan,  nan,  nan,  nan,  31.,  31.,  31.,  31.,
         31.,  31., 160.],
       [ nan,  nan,  nan,  nan,  nan,  nan,  nan,  nan,  nan,  nan,  nan,
         nan,  nan, 160.]])
Coordinates:
  * latitude   (latitude) float32 60.0 59.0 58.0 57.0 ... 53.0 52.0 51.0 50.0
  * longitude  (longitude) float32 -11.0 -10.0 -9.0 -8.0 ... -1.0 0.0 1.0 2.0
```

[9]:
```
<matplotlib.collections.QuadMesh at 0x7f877194e8d0>
```



And now we can extract the UK averaged precipitation within SEAS5 by using the mask index of the UK: `where(SEAS5_mask == UK_index)`. So we need to find the index of one of the 241 abbreviations. In this case for the UK use 'GB'. Additionally, if you can't find a country, use `countries.regions` to get the full names of the countries.

[10]:
```
countries.abbrevs.index('GB')
```

[10]:
```
31
```

To select the UK average, we select SEAS5 precipitation (tprate), select the gridcells that are within the UK and take the mean over those gridcells. This results in a dataset of February precipitation for 35 years (1981-2016), with 5 leadtimes and 25 ensemble members.

[11]:
```
SEAS5_UK = (SEAS5['tprate']
            .where(SEAS5_mask == 31)
            .mean(dim=['latitude', 'longitude']))
SEAS5_UK
```

[11]:
```
<xarray.DataArray 'tprate' (leadtime: 5, time: 35, number: 25)>
array([[[1.7730116 , 1.9548205 , 3.7803986 , ..., 2.9277864 ,
         1.8929143 , 2.446378  ],
        [3.040877  , 1.8855734 , 4.2009687 , ..., 2.8347836 ,
         4.1132317 , 2.4391398 ],
        [3.556001  , 3.6879914 , 3.184576  , ..., 5.116828  ,
         3.8872097 , 2.6074293 ],
        ...,
        [2.9086895 , 3.7759151 , 4.660708  , ..., 1.6525848 ,
         3.1059399 , 1.6731865 ],
        [2.8330274 , 3.6866314 , 2.634609  , ..., 3.3976977 ,
         4.993076  , 3.6399577 ],
        [2.2500532 , 2.510962  , 3.7602315 , ..., 3.8013346 ,
         1.2315732 , 3.6035924 ]],

       [[1.0868028 , 1.5332695 , 3.2461395 , ..., 1.5274822 ,
         2.7315547 , 1.0240313 ],
        [0.99898285, 2.9119303 , 2.1601522 , ..., 3.3752463 ,
         2.8715765 , 4.393921  ],
        [3.6581304 , 3.5088263 , 2.0143754 , ..., 3.9588785 ,
         3.795881  , 3.179954  ],
```

(continues on next page)

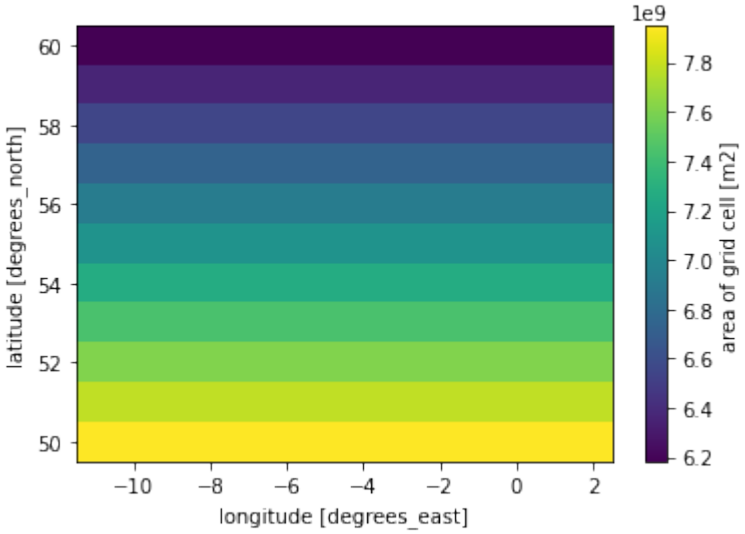
(continued from previous page)

```
        ...,
        [2.5120296 , 3.568533  , 3.4690757 , ..., 2.6206532 ,
         3.953479  , 3.789593  ],
        [1.9866585 , 3.670435  , 2.7552466 , ..., 3.5482445 ,
         1.9896345 , 3.9373026 ],
        [1.9513754 , 2.0166924 , 2.9413762 , ..., 3.3549297 ,
         3.0631557 , 1.915903  ]],

       [[2.0119116 , 2.4435556 , 1.3927166 , ..., 1.8064059 ,
         1.8338976 , 2.4649143 ],
        [2.62523   , 3.6218376 , 3.003645  , ..., 3.3206403 ,
         2.5519    , 3.110555  ],
        [4.071685  , 2.6880858 , 3.8181992 , ..., 1.9033648 ,
         1.8840973 , 4.449085  ],
        ...,
        [2.8379328 , 2.6923704 , 1.82504   , ..., 2.8684394 ,
         3.1954165 , 3.445909  ],
        [2.072118  , 2.0628428 , 2.8790975 , ..., 2.697285  ,
         3.3291562 , 2.9145727 ],
        [3.535856  , 3.6515677 , 2.1052096 , ..., 3.2938933 ,
         3.1968002 , 2.2115657 ]],

       [[2.9105136 , 3.6938024 , 1.1343408 , ..., 1.1984391 ,
         2.9722617 , 2.1267796 ],
        [4.02007   , 1.8249133 , 3.099     , ..., 3.4654658 ,
         2.7237208 , 3.5907636 ],
        [3.1248841 , 2.219241  , 3.6903172 , ..., 0.8401702 ,
         4.017805  , 3.476175  ],
        ...,
        [2.8703861 , 5.5576715 , 1.7174771 , ..., 2.1288645 ,
         1.8581603 , 2.2952495 ],
        [4.179039  , 3.737323  , 2.880745  , ..., 1.9207952 ,
         0.6607291 , 2.8571692 ],
        [1.5515772 , 3.0461147 , 2.0624359 , ..., 2.5376263 ,
         4.744201  , 2.0716472 ]],

       [[3.1285267 , 3.269652  , 2.5995293 , ..., 4.347241  ,
         1.3022097 , 1.4634563 ],
        [2.262867  , 3.3503478 , 2.4287066 , ..., 2.593644  ,
         4.251352  , 3.1974325 ],
        [4.0569496 , 2.156282  , 1.781804  , ..., 4.4861846 ,
         1.8805121 , 4.268968  ],
        ...,
        [1.843329  , 3.1374288 , 2.516779  , ..., 1.3068637 ,
         2.4419744 , 2.2696178 ],
        [4.635685  , 4.429196  , 2.4575038 , ..., 4.441483  ,
         1.3038206 , 2.6521633 ],
        [3.1435733 , 2.614458  , 2.1784708 , ..., 3.662669  ,
         1.6688719 , 1.9709551 ]]], dtype=float32)
Coordinates:
  * time      (time) datetime64[ns] 1982-02-01 1983-02-01 ... 2016-02-01
  * number    (number) int32 0 1 2 3 4 5 6 7 8 9 ... 16 17 18 19 20 21 22 23 24
  * leadtime  (leadtime) int64 2 3 4 5 6
```

However, xarray does not take into account the area of the gridcells in taking the average. Therefore, we have to calculate the area-weighted mean of the gridcells. To calculate the area of each gridcell, I use cdo `cdo gridarea infile outfile`. Here I load the generated file:

[12]:
```
Gridarea_SEAS5 = xr.open_dataset('../UK_example/Gridarea_SEAS5.nc')
Gridarea_SEAS5['cell_area'].plot()
```

[12]:
```
<matplotlib.collections.QuadMesh at 0x7f8771a14bd0>
```



[13]:
```
SEAS5_UK_weighted = (SEAS5['tprate']
                     .where(SEAS5_mask == 31)
                     .weighted(Gridarea_SEAS5['cell_area'])
                     .mean(dim=['latitude', 'longitude'])
                    )
SEAS5_UK_weighted
```

[13]:
```
<xarray.DataArray (leadtime: 5, time: 35, number: 25)>
array([[[1.74715784, 1.91625164, 3.74246331, ..., 2.9025497 ,
         1.87161458, 2.42337871],
        [3.01557164, 1.86355946, 4.23964218, ..., 2.82348107,
         4.08311346, 2.45320866],
        [3.45037457, 3.67373672, 3.19124952, ..., 5.10772697,
         3.83928173, 2.59254324],
        ...,
        [2.91576568, 3.76689869, 4.64815469, ..., 1.61583385,
         3.07634248, 1.633837  ],
        [2.8217756 , 3.61588493, 2.61442489, ..., 3.36894025,
         5.01010622, 3.58215465],
        [2.21954162, 2.46237273, 3.71758684, ..., 3.72850729,
         1.21044478, 3.5506569 ]],

       [[1.08925258, 1.502868  , 3.23383862, ..., 1.49745391,
         2.74434639, 0.98857514],
        [0.96385251, 2.9144073 , 2.14176199, ..., 3.39404417,
         2.837949  , 4.39105086],
        [3.64189637, 3.44186084, 1.96817031, ..., 3.90560029,
         3.78419096, 3.13070979],
        ...,
        [2.46681904, 3.55272741, 3.41184678, ..., 2.58295751,
         3.84194729, 3.71967185],
        [1.96088291, 3.64719353, 2.73561159, ..., 3.56311814,
         2.00379361, 3.89548019],
        [1.89263296, 1.99179138, 2.91457733, ..., 3.37011609,
```

(continues on next page)

(continued from previous page)

```
         3.05092884, 1.89736692]],

       [[1.94362083, 2.4160058 , 1.36431312, ..., 1.78378666,
         1.82209387, 2.42526171],
        [2.57294707, 3.55756557, 2.96458594, ..., 3.26071746,
         2.60226357, 3.13573254],
        [4.13926899, 2.61380816, 3.76440713, ..., 1.87333769,
         1.82042494, 4.46818308],
        ...,
        [2.81221309, 2.69631474, 1.80062933, ..., 2.86429728,
         3.13303958, 3.44787769],
        [2.07369663, 2.03958281, 2.81885547, ..., 2.65322161,
         3.32297921, 2.83923239],
        [3.49884241, 3.63866711, 2.03510114, ..., 3.24500072,
         3.15147085, 2.16834782]],

       [[2.83876303, 3.61651907, 1.0950032 , ..., 1.17176117,
         2.91180608, 2.1047135 ],
        [3.95351432, 1.78778573, 3.08959013, ..., 3.43481603,
         2.72829325, 3.6158294 ],
        [3.13152664, 2.19419128, 3.64975772, ..., 0.83938823,
         4.01702257, 3.46619512],
        ...,
        [2.90135673, 5.59148292, 1.70016201, ..., 2.08532097,
         1.84797942, 2.23519466],
        [4.19777172, 3.74702108, 2.8748067 , ..., 1.88071816,
         0.62889528, 2.79973163],
        [1.51194718, 3.07317605, 2.05956574, ..., 2.53950207,
         4.70096128, 2.04542173]],

       [[3.15063505, 3.23490175, 2.60923731, ..., 4.27225421,
         1.29816875, 1.42799228],
        [2.21017692, 3.32458317, 2.3819878 , ..., 2.53619218,
         4.24655687, 3.21280586],
        [4.07655988, 2.07606666, 1.75961194, ..., 4.47653645,
         1.81676988, 4.21655002],
        ...,
        [1.82563235, 3.10218576, 2.46347745, ..., 1.28711195,
         2.36333743, 2.26540939],
        [4.63821163, 4.43540009, 2.37741808, ..., 4.44346938,
         1.27125796, 2.63467098],
        [3.12914205, 2.58186504, 2.11029128, ..., 3.6635387 ,
         1.68374372, 1.9288108 ]]])
Coordinates:
  * time      (time) datetime64[ns] 1982-02-01 1983-02-01 ... 2016-02-01
  * number    (number) int32 0 1 2 3 4 5 6 7 8 9 ... 16 17 18 19 20 21 22 23 24
  * leadtime  (leadtime) int64 2 3 4 5 6
```

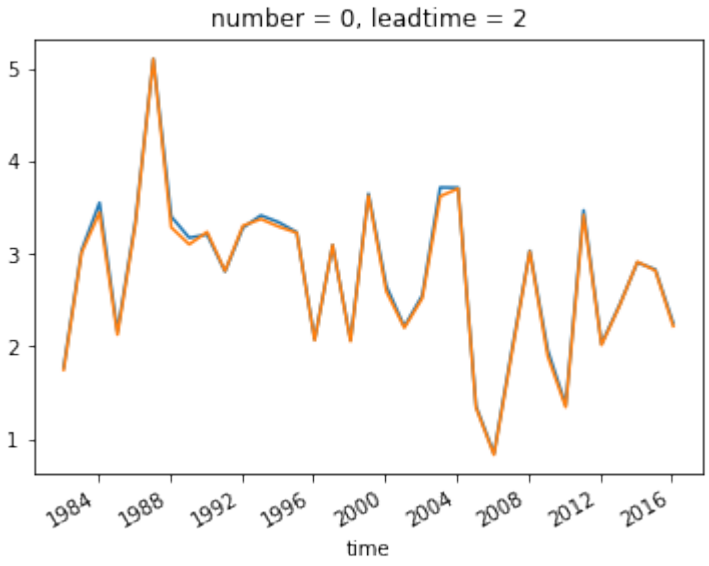
What is the difference between the weighted and non-weighted average?

I plot the UK average for ensemble member 0 and leadtime 2

[14]:
```
SEAS5_UK.sel(leadtime=2,number=0).plot()
SEAS5_UK_weighted.sel(leadtime=2,number=0).plot()
```

[14]:
```
[<matplotlib.lines.Line2D at 0x7f87718f8b50>]
```

[14]:
```
[<matplotlib.lines.Line2D at 0x7f87718e3ad0>]
```



1.11.4 Upscale

For EOBS, we want to upscale the dataset to the SEAS5 grid. We use the function `regridder(ds_in,ds_out, function)`, see the docs. We have to rename the lat lon dimensions so the function can read them.

We use bilinear interpolation first (i.e. function = 'bilinear'), because of its ease in implementation. However, the use of conservative areal average (function = 'conservative') for upscaling is preferred (Kopparla, 2013).

[8]:
```
regridder = xe.Regridder(EOBS.rename({'longitude': 'lon', 'latitude': 'lat'}), SEAS5.
↪rename({'longitude': 'lon', 'latitude': 'lat'}),'bilinear')
```

```
Overwrite existing file: bilinear_201x464_11x14.nc
 You can set reuse_weights=True to save computing time.
```

Now that we have the regridder, we can apply the regridder to our EOBS dataarray:

[27]:
```
EOBS_upscaled = regridder(EOBS)
EOBS_upscaled
```

```
using dimensions ('latitude', 'longitude') from data variable rr as the horizontal␣
↪dimensions for this dataset.
```

[27]:
```
<xarray.Dataset>
Dimensions:  (lat: 11, lon: 14, time: 71)
Coordinates:
  * time     (time) datetime64[ns] 1950-02-28 1951-02-28 ... 2020-02-29
  * lon      (lon) float32 -11.0 -10.0 -9.0 -8.0 -7.0 ... -2.0 -1.0 0.0 1.0 2.0
  * lat      (lat) float32 60.0 59.0 58.0 57.0 56.0 ... 54.0 53.0 52.0 51.0 50.0
Data variables:
    rr       (time, lat, lon) float64 nan nan nan nan nan ... nan nan nan 4.243
Attributes:
    regrid_method:  bilinear
```

And set the latlon dimension names back to their long name. This is so both SEAS5 and EOBS have the same latlon dimension names which is necessary when using the same mask.

[28]:
```
EOBS_upscaled = EOBS_upscaled.rename({'lon' : 'longitude', 'lat' : 'latitude'})
```

1.11.5 Illustrate the SEAS5 and EOBS masks for the UK

Here I plot the masked mean SEAS5 and upscaled EOBS precipitation. This shows that upscaled EOBS does not contain data for all gridcells within the UK mask (the difference between SEAS5 gridcells and EOBS gridcells with data). We can apply an additional mask for SEAS5 that masks the grid cells that do not contain data in EOBS.

[30]:
```
fig, axs = plt.subplots(1, 2, subplot_kw={'projection': ccrs.OSGB()})

SEAS5['tprate'].where(SEAS5_mask == 31).mean(
    dim=['time', 'leadtime', 'number']).plot(
    transform=ccrs.PlateCarree(),
    vmin=0,
    vmax=8,
    cmap=plt.cm.Blues,
    ax=axs[0])

EOBS_upscaled['rr'].where(SEAS5_mask == 31).mean(dim='time').plot(
    transform=ccrs.PlateCarree(),
    vmin=0,
    vmax=8,
    cmap=plt.cm.Blues,
    ax=axs[1])

for ax in axs.flat:
    ax.coastlines(resolution='10m')

axs[0].set_title('SEAS5')
axs[1].set_title('EOBS')
```

```
/soge-home/users/cenv0732/.conda/envs/upscale/lib/python3.7/site-packages/xarray/core/
↪nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

[30]:
```
<matplotlib.collections.QuadMesh at 0x7fc1ea756c50>
```

```
/soge-home/users/cenv0732/.conda/envs/upscale/lib/python3.7/site-packages/xarray/core/
↪nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

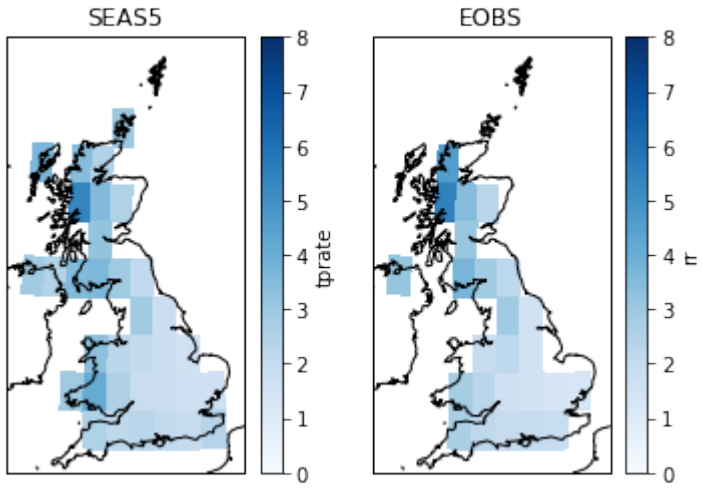
[30]:
```
<matplotlib.collections.QuadMesh at 0x7fc1ea71e650>
```

[30]:
```
<cartopy.mpl.feature_artist.FeatureArtist at 0x7fc1ea6d8f10>
```

[30]:
```
<cartopy.mpl.feature_artist.FeatureArtist at 0x7fc1ea7b9590>
```

[30]:
```
Text(0.5, 1.0, 'SEAS5')
```

[30]:
```
Text(0.5, 1.0, 'EOBS')
```



The additional mask of SEAS5 is where EOBS is not null:

[32]:
```
fig, axs = plt.subplots(1, 2, subplot_kw={'projection': ccrs.OSGB()})

(SEAS5['tprate']
 .where(SEAS5_mask == 31)
 .where(EOBS_upscaled['rr'].sel(time='1950').squeeze('time').notnull()) ## mask␣
↪values that are nan in EOBS
 .mean(dim=['time', 'leadtime', 'number'])
 .plot(
    transform=ccrs.PlateCarree(),
    vmin=0,
    vmax=8,
    cmap=plt.cm.Blues,
    ax=axs[0])
)

EOBS_upscaled['rr'].where(SEAS5_mask == 31).mean(dim='time').plot(
    transform=ccrs.PlateCarree(),
    vmin=0,
    vmax=8,
    cmap=plt.cm.Blues,
    ax=axs[1])

for ax in axs.flat:
    ax.coastlines(resolution='10m')

axs[0].set_title('SEAS5')
axs[1].set_title('EOBS')
```

```
/soge-home/users/cenv0732/.conda/envs/upscale/lib/python3.7/site-packages/xarray/core/
↪nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

[32]:
```
<matplotlib.collections.QuadMesh at 0x7fc1ea579590>
```

```
/soge-home/users/cenv0732/.conda/envs/upscale/lib/python3.7/site-packages/xarray/core/
↪nanops.py:142: RuntimeWarning: Mean of empty slice
  return np.nanmean(a, axis=axis, dtype=dtype)
```

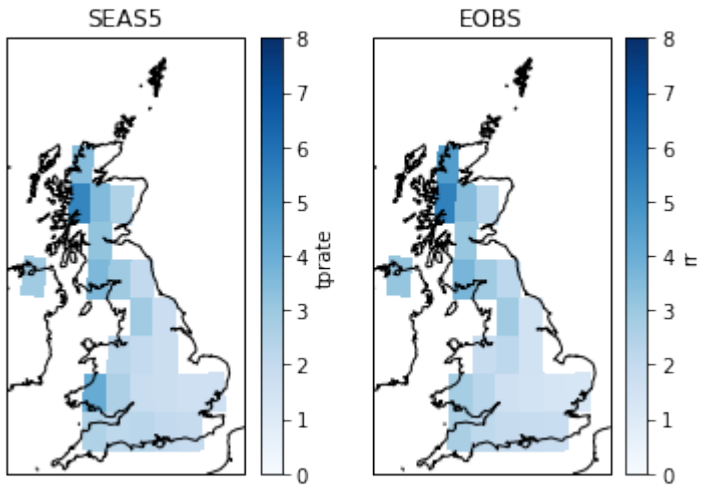
```
[32]: <matplotlib.collections.QuadMesh at 0x7fc1ea56f750>
```

```
[32]: <cartopy.mpl.feature_artist.FeatureArtist at 0x7fc1ea4fe450>
```

```
[32]: <cartopy.mpl.feature_artist.FeatureArtist at 0x7fc1ea56ffd0>
```

```
[32]: Text(0.5, 1.0, 'SEAS5')
```

```
[32]: Text(0.5, 1.0, 'EOBS')
```



1.11.6 Extract the spatial average

To select the UK average, we select SEAS5 precipitation (tprate), select the gridcells that are within the UK and take the area-weighted mean over those gridcells. This results in a dataset of February precipitation for 35 years (1981-2016), with 5 leadtimes and 25 ensemble members.

```
[38]: SEAS5_UK_weighted = (SEAS5
                          .where(SEAS5_mask == 31)
                          .where(EOBS_upscaled['rr'].sel(time='1950').squeeze('time').
      ↪notnull())
                          .weighted(Gridarea_SEAS5['cell_area'])
                          .mean(dim=['latitude', 'longitude'])
                         )
      SEAS5_UK_weighted
```

```
[38]: <xarray.Dataset>
      Dimensions:    (leadtime: 5, number: 25, time: 35)
      Coordinates:
        * time       (time) datetime64[ns] 1982-02-01 1983-02-01 ... 2016-02-01
        * number     (number) int32 0 1 2 3 4 5 6 7 8 9 ... 16 17 18 19 20 21 22 23 24
        * leadtime   (leadtime) int64 2 3 4 5 6
      Data variables:
          tprate     (leadtime, time, number) float64 1.62 1.803 3.715 ... 1.681 1.769
```

```
[40]: EOBS_UK_weighted = (EOBS_upscaled
                          .where(SEAS5_mask == 31) ## EOBS is now on the SEAS5 grid, so use␣
      ↪the SEAS5 mask and gridcell area
                          .weighted(Gridarea_SEAS5['cell_area'])
                          .mean(dim=['latitude', 'longitude'])
```

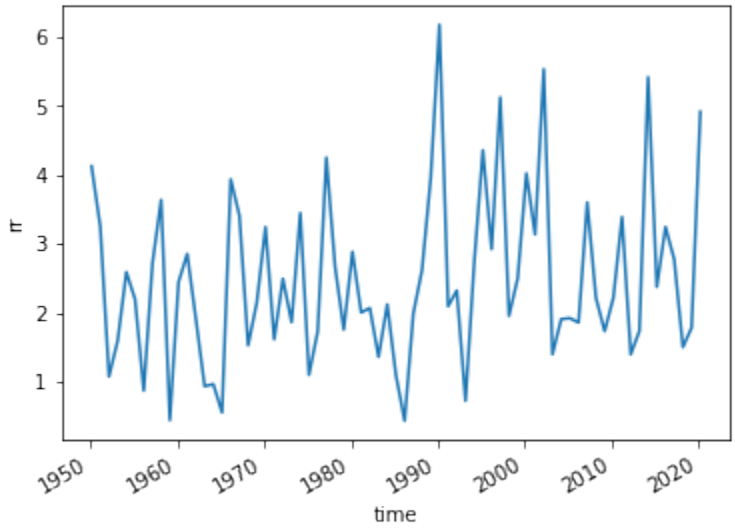
(continues on next page)

(continued from previous page)

```
                  )
EOBS_UK_weighted
EOBS_UK_weighted['rr'].plot()
```

[40]:
```
<xarray.Dataset>
Dimensions:  (time: 71)
Coordinates:
  * time     (time) datetime64[ns] 1950-02-28 1951-02-28 ... 2020-02-29
Data variables:
    rr       (time) float64 4.127 3.251 1.072 1.593 ... 2.774 1.498 1.782 4.92
```

[40]:
```
[<matplotlib.lines.Line2D at 0x7fc1ea2f4850>]
```



1.11.7 Illustrate the SEAS5 and EOBS UK average

And the area-weighted average UK precipitation for SEAS5 and EOBS I plot here. For SEAS5 I plot the range, both min/max and the 2.5/97.5 % percentile of all ensemble members and leadtimes for each year.

[43]:
```
ax = plt.axes()

Quantiles = (SEAS5_UK_weighted['tprate']
             .quantile([0,2.5/100, 0.5, 97.5/100,1],
                       dim=['number','leadtime']
                      )
            )
ax.plot(Quantiles.time, Quantiles.sel(quantile=0.5),
        color='orange',
        label = 'SEAS5 median')
ax.fill_between(Quantiles.time.values, Quantiles.sel(quantile=0.025), Quantiles.
↪sel(quantile=0.975),
                color='orange',
                alpha=0.2,
                label = '95% / min max')
ax.fill_between(Quantiles.time.values, Quantiles.sel(quantile=0), Quantiles.
↪sel(quantile=1),
                color='orange',
```

(continues on next page)

(continued from previous page)

```
                alpha=0.2)

EOBS_UK_weighted['rr'].plot(ax=ax,
                            x='time',
                            label = 'E-OBS')
plt.legend(loc = 'lower left',
           ncol=2 ) #loc = (0.1, 0) upper left
```

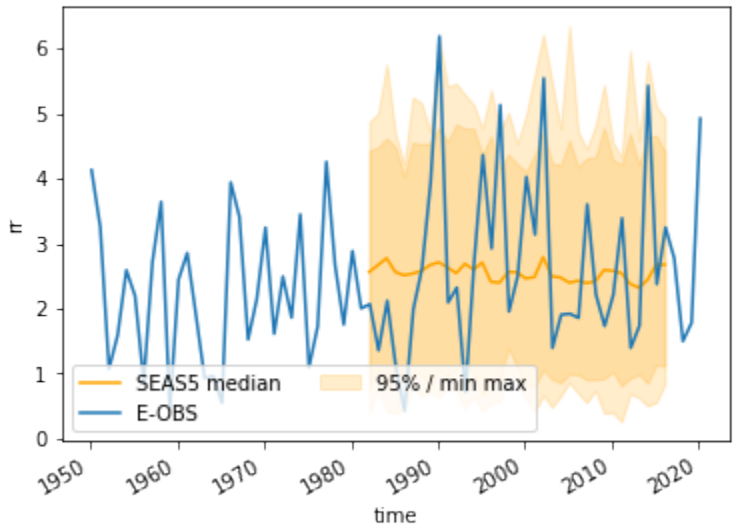
```
[43]: [<matplotlib.lines.Line2D at 0x7fc1ea2cef10>]
```

```
[43]: <matplotlib.collections.PolyCollection at 0x7fc1ea39f250>
```

```
[43]: <matplotlib.collections.PolyCollection at 0x7fc1ea77fed0>
```

```
[43]: [<matplotlib.lines.Line2D at 0x7fc1ea449310>]
```

```
[43]: <matplotlib.legend.Legend at 0x7fc1ea56fc90>
```



1.11.8 And save the UK weighted average datasets

```
[45]: SEAS5_UK_weighted.to_netcdf('Data/SEAS5_UK_weighted_masked.nc')
      SEAS5_UK_weighted.to_dataframe().to_csv('Data/SEAS5_UK_weighted_masked.csv')
      EOBS_UK_weighted.to_netcdf('Data/EOBS_UK_weighted_upscaled.nc') ## save as netcdf
      EOBS_UK_weighted.to_dataframe().to_csv('Data/EOBS_UK_weighted_upscaled.csv') ## and␣
      ↪save as csv.
```

```
[46]: SEAS5_UK_weighted.close()
      EOBS_UK_weighted.close()
```

1.11.9 Other methods

There are many different sources and methods available for extracting areal-averages from shapefiles. Here I have used shapely / masking in xarray. Something that lacks with this method is the weighted extraction from a shapefile, that is more precise on the boundaries. In R, raster:extract can use the percentage of the area that falls within the country for each grid cell to use as weight in averaging. For more information on this method, see the EGU 2018 course. For SEAS5, with its coarse resolution, this might make a difference. However, for it's speed and reproducibility, we have chosen to stick to xarray.

We have used xarray where you can apply weights yourself to a dataset and then calculate the weighted mean. Sources I have used: * xarray weighted reductions * Matteo's blog * regionmask package * Arctic weighted average example * area weighted temperature example.

And this pretty awesome colab notebook on seasonal forecasting regrids seasonal forecasts and reanalysis on the same grid before calculating skill scores.

CHAPTER

TWO

LICENSE

All code and example data are available under the open source MIT License.

CHAPTER

THREE

CITATION

When using the code or example data, please cite this project. If any questions may arise, please don't hesitate to get in touch t.kelder@lboro.ac.uk.