Alma Mater Studiorum – Università di Bologna

Dipartimento di Fisica e Astronomia "Augusto Righi"
Laurea Magistrale in Fisica del Sistema Terra

Dynamical Downscaling of Hybrid Seasonal Predictions

Presentata da:
Gabriele Bentivoglio

Relatore:
Dott. Paolo Ruggieri

Correlatore:
Prof.ssa Silvana Di Sabatino

Anno Accademico 2023 - 2024
Appello III

Sommario

Le ondate di calore estive sono aumentate in frequenza negli ultimi decenni in Europa (Rousi et al. 2022), così come il loro impatto sulla salute pubblica e sul sistema produttivo. Prevedere questi eventi su una scala sub-stagionale e stagionale sarebbe cruciale per mitigare le loro conseguenze negative (C. J. White et al. 2017). Questa tesi propone una metodologia che integra *ensemble subsampling* e downscaling dinamico per migliorare la rappresentazione delle temperature estive, concentrandosi sulla città di Bologna e sulle aree rurali circostanti.

L'aspetto principale consiste nella preparazione di un dataset di previsioni stagionali per inizializzare il modello di downscaling. La procedura combina dataset pubblicamente disponibili per garantire una facile replicabilità. Attraverso test di sensibilità, viene identificata una configurazione adeguata del modello WRF per effettuare il downscaling su un sottoinsieme di membri dell'*ensemble* di previsioni stagionali.

Il downscaling dinamico riduce il BIAS e il MAE della temperatura media mensile a due metri in tutte le località considerate, con i maggiori benefici osservati nelle aree urbane. L'introduzione dell'aggiornamento della temperatura superficiale del mare in WRF riduce il BIAS della temperatura sulle aree marine, ma ulteriori correzioni sarebbero necessarie per una rimozione completa.

Viene esplorata anche la performance di semplici correzioni statistiche, evidenziando il potenziale di un approccio ibrido parallelo che combini downscaling dinamico e *Mean and Variance Adjustment.*

Riconoscendo le limitazioni dell'attuale configurazione, vengono delineati i possibili passi futuri. Questi includono l'implementazione di un *process-informed ensemble subsampling* e il miglioramento dell'accoppiamento oceano-atmosfera. Inoltre, vengono presentate le metriche raccomandate per una valutazione più completa dei risultati.

Abstract

Summer heatwaves have been increasing in frequency in the past decades over Europe (Rousi et al. 2022), and so has their impact on public health and the productive system. Predicting these events on a subseasonal-to-seasonal timescale would be crucial to mitigate their negative effects (C. J. White et al. 2017). This thesis proposes a methodology integrating ensemble subsampling and dynamical downscaling to improve the representation of summer temperatures, focusing on the city of Bologna and the neighbouring rural areas.

The main aspect concerns the preparation of a seasonal forecast dataset to initialize the downscaling model. The procedure combines publicly available datasets to ensure easy replicability. Through sensitivity tests, I identify an adequate configuration of the WRF model to downscale a subset of members from the seasonal forecast.

The dynamical downscaling reduces the monthly-averaged two-metre temperature BIAS and MAE across all locations considered, with greater benefit observed in the urban locations. Introducing the sea surface temperature update in WRF reduces the temperature BIAS over marine areas, but further corrections would be needed to address it fully.

The performance of simple statistical corrections is also explored, highlighting the potential of combining the dynamical downscaling with the Mean and Variance Adjustment technique within a hybrid parallel approach.

This work also acknowledges the limitations of the current setup while outlining the possible future steps. These include implementing a process-informed ensemble subsampling and improving the ocean-atmosphere coupling. Additionally, I present the recommended metrics for a more comprehensive evaluation of the results.

Contents

Chapter 1

Introduction

The effect of heatwaves on public health and agriculture can be mitigated with appropriate planning, but this requires the capability to anticipate such events weeks or even months in advance. Focusing on Northern Italy, the Po Valley is a densely populated region with multiple urban areas. Cities are the most susceptible locations to heatwaves, as they can be amplified by the urban fabric because of the Urban Heat Island (UHI) effect, of which a review is presented in Deilami, Kamruzzaman, and Liu 2018. Possega et al. 2022 evaluates the multiscale relationship between heatwaves and UHI for multiple European cities. Rural areas are not exempt from the effects of heatwaves, which can place a significant strain on water management by reducing precipitation and typically increasing irrigation needs, as explained in Cárdenas Belleza, Bierkens, and Van Vliet 2023. This is critical given the intensive agricultural practices in the region.

The seasonal forecasts produced by the ECMWF show a negative BIAS over the Po Valley area across the reference period taken into consideration. Further details are presented in Section 2.1. Removal or reduction of this BIAS may help improve the quality and usefulness of the forecast.

My thesis aim is to explore whether a combination of existing well-established techniques has the potential to improve the representation of surface temperature in the city of Bologna, which lies in the southern part of the Po Valley. More precisely, the goal is to propose a methodology based on dynamical downscaling to improve forecast skill and reliability within the subseasonal-to-seasonal range. The downscaling is preceded by the introduction of an ensemble subsampling, to reduce the total computational cost of the operation. This work is intended as a foundation on which to build upon. The known limitations and proposed next steps are outlined in Section 2.5.

The key part of the methodology presented in this thesis is the construction of the dataset that will be used to initialize the downscaling model. This is done through a combination of freely accessible data from the Copernicus Climate Data Store (CDS) which is then adapted to initialize the downscaling model.

This setup is then tested on a set of locations within and around the city of Bologna, by comparing the results before and after the downscaling against the observational reference provided by the in-situ weather stations and available larger-scale datasets. Alternative simple statistical correction techniques are also introduced.

In the following sections of the introduction, some foundational concepts are briefly discussed before continuing. Next, in Chapter 2, I describe the methodology in place. Here the different sources of data are presented, followed by the models and the simulation setup. Additionally, I explain how the seasonal dataset has been modified starting from the available data to adapt it to the WRF operations. In Chapter 3, I present the results of preliminary sensitivity tests on the WRF model setup, aimed at assessing the downscaling quality and evaluating any significant changes. The same scores are then evaluated again for a set of selected configurations for a longer period. The downscaling of a seasonal forecast member is also checked. The results are shown in Chapter 4, where it is explored the performance of the ensemble mean of the members downscaled using the final setup, together with the alternative statistical correction methods. Ultimately, the most significant outcomes and the future steps are discussed in Section 5.

1.1 The subseasonal-to-seasonal range

The potential benefit of considering the information provided by external forcing and boundary conditions has been clear since the Seventies. Madden 1976 showed that at the mid-latitudes, a significant portion of the total variability can be attributed to them. Long-range predictability was seen as an estimated upper bound of skill, which had not been reached yet.

Among the others, the subseasonal-to-seasonal (S2S) range can benefit from this assessment. S2S is only marginally influenced by the initial condition, while the role of the slowly evolving boundary conditions is more relevant (Shukla 1998). These include sea surface temperature (SST), soil moisture and sea ice cover. These quantities are crucial in the correct delineation of the interaction between the atmosphere and the Earth's surface through heat and moisture fluxes, and anomalies can influence the forecast outcome (Schwitalla et al. 2008).

Among the land attributes, the most impactful is soil moisture (Merryfield et al. 2020). Long-lasting soil moisture anomalies have proven to be determinant in certain areas, with opposing extremes generating skill in different regions. It is also suggested by Ferranti and Viterbo 2006 that perturbing the initial soil moisture may help to handle the associated uncertainties, a further argument in favour of ensemble approaches. There is a significant impact on air temperature forecast skill in those areas where the underlying ground observation network is reliable, as found by Koster, Mahanama, Yamada, Balsamo, Berg, Boisserie, Dirmeyer, Doblas-Reyes, et al. 2010 for North America and later confirmed in Koster, Mahanama, Yamada, Balsamo, Berg, Boisserie, Dirmeyer,

F. J. Doblas-Reyes, et al. 2011, which extended the same analysis to a global scale. More in general, this highlights the importance of a realistic land surface initialization. Van Den Hurk et al. 2012 focuses instead on the European region, where it is shown there is still a quantifiable benefit in subsampling members with similar soil moisture conditions, or with extreme soil moisture conditions. However, temperature forecasts show lower skill improvements compared to the US. This may be due to the influence of large and remote Atlantic air masses on air temperature and precipitation.

Other sources of predictability are processes connected to climate variability, like the North-Atlantic Oscillation (NAO) and the Madden-Julian Oscillation (MJO), or the interaction with the stratosphere, as summarized in Mariotti et al. 2020; Merryfield et al. 2020; Meehl et al. 2021. Given their longer timescale, they impart memory into the system and can be exploited to disclose hidden predictability. An overview of the main predictability sources depending on the time range is shown in Figure 1.1.

To take advantage of these sources of information, a larger number of ensemble members may be preferable, as it helps to better predict NAO and MJO patterns. Improved subseasonal forecast skill is observed when tailoring the ensemble generation approach tailored for MJO prediction (Kim, Vitart, and Waliser 2018). The additional computational cost entailed should always be considered. Another area of improvement is the coupling with the ocean, as it would allow to improve the use of the information coming from SST and sea-ice. However, it is required a deeper understanding of the mechanisms involved and their role in the climate system.

1.2 Ensemble forecasts and subsampling

Global models are currently the main tool to produce seasonal climate forecasts (Manzanas, J. Gutiérrez, et al. 2018). Operative forecasts are relatively a novelty in this range and many aspects of the system design have undergone significant changes in the past few years, with a major emphasis on the representation of initial conditions and model physics uncertainty (Merryfield et al. 2020). Beyond the operational forecast systems that have been recently introduced by ECMWF, NCEP, UK Met Office and other major agencies, the datasets from *S2S* and *SubX MME* projects have also been available for research purposes.

Given the chaotic nature of the atmosphere, as anticipated in Section 1.1, an advantage is given by an ensemble approach where instead of a single global model, different runs are considered together. Each one of them is initialized with a perturbed initial condition and this allows to better capture the variability of the system. The ensemble approach is necessary to better capture the variability of the system since the current models' skill is limited. For instance, the SubX project models show skill for temperature and precipitation 3 weeks ahead of time only in specific regions, as shown by Pegion et al. 2019. The ECMWF system has been shown in De Andrade, Coelho, and Cavalcanti

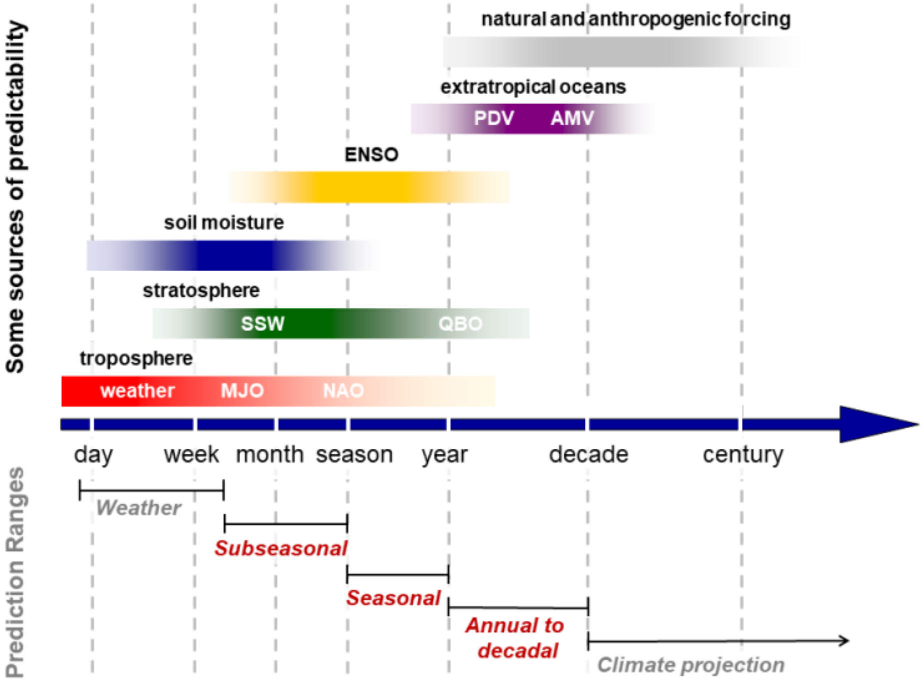

Figure 1.1: Main sources of predictability depending on the time range. The S2S range is mostly affected by the state of soil moisture and climate variability processes. Information can also be inferred from stratospheric processes, like the Sudden Stratospheric Warming (SSW). The Figure is taken from Mariotti et al. 2020.

2019 to be among the top-scoring models, but the scores still decrease with increasing lead time.

One key area of focus is now the identification of smaller subsets within the ensemble to increase the skill of the forecast system. This has been done in multiple ways. In Albers and Newman 2019 a Linear Inverse Model (LIM) is used to establish a priori the expected forecast skill and subsequently identify the subset within the ECMWF system that shows higher skill. The selection ultimately comes down to the specific location and season of interest, but a LIM can be a useful tool to detect the so-called forecasts of opportunity, which increased prediction ability is a consequence of the influence of phenomena which yield memory, and thus information. In these cases, a predictable signal is dominant over the unpredictable noise in the system (Breeden et al. 2022). In Mori et al. 2021 a subsampling of the ensemble forecast system is done by taking a small number of members which spread is a good representation of the ensemble spread for SST, an essential variable for the area being studied in the article.

The subsampling can also be process-informed when it is based on the ability of the ensemble member to effectively represent one or multiple chosen predictors (Dobrynin et al. 2018; Kowal et al. 2024). A predictor is an indicator having a well-understood physical connection with a key variable or with a climate index, for example, the ones describing the phase of a climate variability process. It shows a strong statistical correlation with said quantity in the region of interest. Finding relevant predictors can be a challenge, as it can be a time-consuming and computationally expensive process. With the recent development of Machine Learning (ML) and Artificial Intelligence (AI) it is getting easier to build more advanced statistical relationships for a certain predictand. One main challenge is to ensure the statistical connection found represents a physical process, which could be unknown. AghaKouchak et al. 2022 recently proposed a bottom-up approach that can help with the identification. The use of ML and AI must be especially cautious when dealing with extreme events, as they are likely not part of the training set. This can be exacerbated by global warming trends, and it may cause a significant degradation in performance, as explained by Miloshevich et al. 2022 and Materia et al. 2024.

1.2.1 The benefits of dynamical downscaling

Given the coarse resolution of global models, they are unable to provide useful information on a regional scale without further elaboration. The raw approximation of the topography impacts the ability of the model to correctly represent local features. This leads to low prediction skill and extensive bias over some regions.

One of the most widely explored solutions is dynamical downscaling, which consists in nesting the global model with a local area model, which has a finer grid and is configured for the specific region of interest. This is an expensive approach, where the local area model is driven by the global one through lateral boundary conditions and initial conditions. Nevertheless, it has been shown to improve the quality of forecasts in areas with complex topography (Schwitalla et al. 2008) and bodies of water (Lauwaet et al. 2012). An appropriate choice for the set of land use categories can further increase the quality of the forecast, as assessed through downscaling with the WRF model in López-Espinoza et al. 2020. Another setting where it has proven an advantage in terms of bias and skill is in regional subseasonal-to-seasonal precipitation forecasts. This has been shown in Pal et al. 2019, where a downscaling is performed with a regional convection-permitting model while considering periods with mostly convective precipitation. This led to a reduction in mean and extreme summer precipitation bias. The bias in the regional model output is different from the one from global models, which was largely due to the misrepresentation of topography, whereas in this case it is more process-related, as explained in Manzanas, Lucero, et al. 2018. The downscaling is also critical for risk analysis, as high-resolution data is required for optimal results (Brogno et al. 2023).

One important aspect to consider is the coexistence of different sources of uncertainty. Other than the global forecasting system, the regional model is often initialized with a reanalysis, to assess to which extent the improvements are imputable to the improved land initialization and forcing. This clarifies to what extent the error is due to the global model providing the initial condition and which part is instead given by the regional model integration (Gulilat Tefera Diro 2016). There are then techniques to mitigate both the uncertainty on the initial condition, through perturbations, and of the model error, through stochastic physical schemes (Anderson et al. 2007). To represent the uncertainty on the initial and boundary conditions it can be helpful to introduce an ensemble approach, ensuring that the associated spread accurately reflects that of the fields, as suggested in Mori et al. 2021.

1.3 Heatwaves and their predictability

A heatwave is an insistent condition of anomalously higher temperature in a certain region. Definitions are ambiguous and there have been attempts to standardize using percentile thresholds and restrictions on duration. The first range of proposed solutions can be traced back to Robinson 2001. In the more recent reviews by Perkins and Alexander 2013 and Domeisen et al. 2022 the question is addressed again and while there is agreement on the general characteristics, the definition remains application-specific. One interesting option is the *Heatwave Magnitude Index*, introduced by Russo et al. 2014 to account for both heatwave amplitude and duration, while also considering the context of the local climate in which the extreme event takes place. Building on this concept, the recent study by Prodhomme et al. 2022 introduces the *Heatwave Propensity*, which is then used to assess the ability of a model to predict the predisposition of a season to heatwave occurrence over Europe.

There is evidence of an increase in the latest decades of the global coverage of heatwaves and their magnitude, as indicated in Zampieri et al. 2016. Additionally, the Northern Hemisphere has experienced a higher number of extreme heat events, especially over Europe, which Rousi et al. 2022 identified as a *heatwave hotspot*. This boosted the interest in predicting such events with skilful forecasts, given their impact on public health (Campbell et al. 2018), agriculture (Ribeiro et al. 2020) and how they can increase wildfire risk (Libonati et al. 2022) and reduce water availability (Zampieri et al. 2016).

There are multiple valuable sources of predictability in the subseasonal-to-seasonal forecast of such events. As mentioned in Section 1.1, land-atmosphere interaction plays an important part. There exists a coupling between lack of spring precipitation and extreme summer temperatures (Fischer et al. 2007), which is likely due to a feedback effect with soil moisture (Seneviratne et al. 2010). A statistical linkage has also been found between upper-tropospheric transient Rossby wave packets and surface temperature extremes, especially at mid-latitudes, even though this interaction seems to be case-dependent (Fragkoulidis et al. 2018). While this is not the case for heatwaves over Europe, a relationship has been found even with the regional stalling of the jet stream meanders (Röthlisberger, Stephan Pfahl, and Martius 2016).

Summer heatwaves tend to be the more predictable on a subseasonal range. The extreme heatwaves can often manifest in ensemble forecasts as warm anomalies at lead times even longer than a month Domeisen et al. 2022. While these extreme events tend to be more predictable with respect to the general skill for a temperature forecast, this is certainly region-sensitive and higher skill is only observed in the first two weeks of lead time, as shown in Wulff and Domeisen 2019.

Another aspect of interest is how the characteristics of a single heatwave influence its predictability. In a study conducted over the European region by Pyrina and Domeisen 2023, it is shown that the intensity of the heatwave conditions the predictability of its intensity. Whether it favours it or not, it depends on the region and lead week. Concerning the heatwave onset, more intense heatwaves are associated with a higher predictability in the first lead week. The number of correctly predicted onsets is very low beyond that threshold for any intensity. The predictability of the heatwave duration is very low as well beyond the first lead week, but Lavaysse et al. 2019 suggests that selecting fewer ensemble members increases the number of correctly represented events.

Heatwaves in Europe

When studying heatwaves in the European region, it is of unequivocal importance the influence of low-frequency climate variability in the Euro-Atlantic region. Summer NAO, while smaller in amplitude compared to its winter equivalent, is still able to influence the climate pattern of northwestern Europe. The second dominant mode of summer variability in the area is the Summer East Atlantic pattern, and it can be associated to the dynamics of summer heatwaves in Europe as well (Wulff, Greatbatch, et al. 2017). These are, in turn, weakly influenced by ENSO (Folland et al. 2009) and MJO teleconnections (Merryfield et al. 2020).

The connection between spring precipitation and summer temperature anomalies anticipated in Section 1.3 is present in Europe too. Indeed, Quesada et al. 2012 shows the role of a particularly wet spring season in inhibiting summer heatwaves over Southern Europe. On the other hand, a dry season can lead to decreased predictability, since in that case there seems to be an increased sensitivity to the weather regimes.

The role of atmospheric blockings and subtropical ridges

The occurrence of atmospheric blocking, which obstructs the prevailing flows and is usually associated with a high-pressure area, is often co-located with a heat extreme, especially at higher latitudes, as indicated by S. Pfahl and Wernli 2012. One reason can be the establishment of a large-scale subsidence, which yields the absence of clouds and increases the radiative warming on summer days, as explained by Kautz et al. 2022. While a strong correlation between blocking and heatwaves is found in Northern Europe, there seems to be a strong anti-correlation in Southern Europe in all seasons (Brunner et al. 2018). There are, in fact, different categories of blocking systems and depending on their configuration they are prevalent in certain regions and latitudes and lead to distinct effects. A general review of blocking climatology is available in Lupo 2021, while Kautz et al. 2022 provides a review of the effects of different kinds of blocking systems in terms of surface weather extremes.

Beyond enhancing the radiative heating of the surface, atmospheric blockings can favour the occurrence of high-temperature anomalies through heat accumulation due to increased large-scale warm air advection and vertical advection, as shown in Miralles et al. 2014. It has been noted in Sousa et al. 2018 that the latter, and its associated adiabatic heating, only plays a secondary role and is more relevant during winter and as we move closer to the centre of the high-pressure area.

During 2023 summer, there have been three major anomalously warm periods over Italy, namely in the second half of June, around the middle of July and in the second half of August. All of them can be associated on a synoptic scale with an amplified ridge, which dynamics and effects on the weather differ from the ones of the atmospheric blockings seen further north (Sousa et al. 2018). This structure is typical of lower latitudes like those of Southern Europe. It is a key driver in the development of temperature extremes, with diabatic heating contributing more to positive summer heat anomalies compared to horizontal advection (Sousa et al. 2018). Ultimately, it is important to consider the challenges that remain in the correct simulation of these large-scale persistent patterns, as remarked in Domeisen et al. 2022. Their predictions can carry a high degree of uncertainty.

Chapter 2

Methodology

In this Chapter, I present the data sources that are being used, and I also provide the details of the methodology to build the dataset initializing the downscaling model. An overview is presented in Figure 2.1, which highlights the most relevant phenomena and briefly anticipates how the data is combined to create the initial and boundary conditions. Further details and instructions that allow to replicate the operations are described in Section 2.2.

The data is described in Section 2.1 and consists of ground station instantaneous measurements, and the data from ERA5 and E-OBS datasets, which details will be provided in Sections 2.1 and 2.1 respectively.

Next, in Section 2.3, the models in use are described, together with their tested configurations. Subsequently, in Section 2.7, I provide a brief review of the recommended quality metrics for testing the setup and analysing the resulting data.

2.1 Data

In this work, I focus on surface temperature and relative humidity for the summer in the Emilia-Romagna region, Northern Italy. The data comprises observations from a set of local weather stations, ERA5 reanalysis data and the seasonal forecasts of the ECMWF ensemble forecast system.

Stations data

As a reference when comparing different simulations, the data from five weather stations located around and within the city of Bologna are considered. The stations are Bologna Urbana (also referred to by the abbreviation BOU), Bologna Idrografico (BOI), Bologna Asinelli (BOA), Mezzolara (MEZ), Sant'Agata Bolognese (STG) and San Pietro Capofiume (SPT). The data is provided by ARPAE, the local regional environment agency maintaining the stations, through their $D3xt3r$ portal. The data is acquired every half

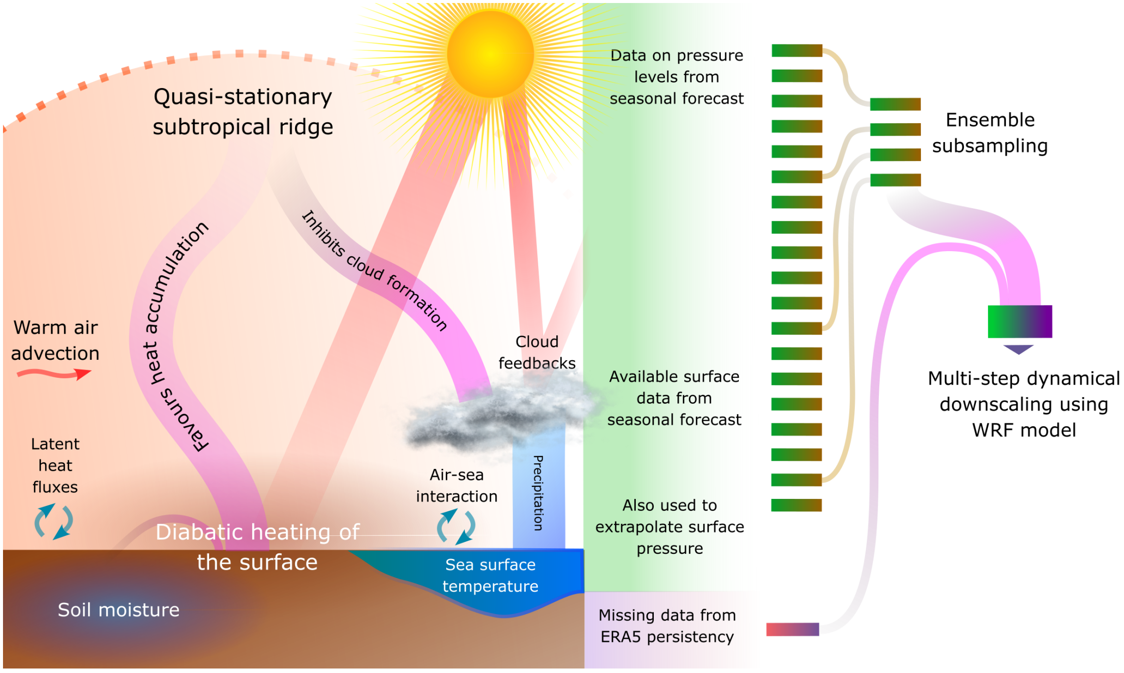
Quasi-stationary subtropical ridge
Inhibits cloud formation
Favours heat accumulation
Warm air advection
Cloud feedbacks
Latent heat fluxes
Air-sea interaction
Precipitation
Diabatic heating of the surface
Sea surface temperature
Soil moisture
Data on pressure levels from seasonal forecast
Available surface data from seasonal forecast
Also used to extrapolate surface pressure
Missing data from ERA5 persistency
Ensemble subsampling
Multi-step dynamical downscaling using WRF model

Figure 2.1: On the left-hand side, this figure presents the main processes involved in the onset and persistence of a heatwave event in the Mediterranean area. This is an adaptation from Domeisen et al. 2022. Being a subtropical region, the associated synoptic-scale pattern is an extended subtropical ridge. An important role is also played by the different feedbacks at the surface. Further details are presented in Section 1.3. The main steps of the methodology are outlined on the right-hand side of the Figure. They represent how the data from open-source services is combined to build the dataset, which then constitutes the initial and boundary conditions for the downscaling model. The green rectangles represent the members of the global seasonal forecast, while the purple one represents the ERA5 data which is re-gridded and complements the dataset. Further details on what data is available in each dataset are described in Section 2.4 and summarized in Table 2.1.

an hour, except for the San Pietro Capofiume station which records a value every 15 minutes. Regular soundings are also available in correspondence with this station, with profiles measured every 12 hours, at 00:00 GMT and 12:00 GMT. Bologna Urbana, Bologna Idrografico and Bologna Asinelli are located within the urban area of Bologna, while the other three stations are placed in the flat rural area around the city to its north. Specifically, I am using two-metre temperature and relative humidity instantaneous data for all the stations among those cited where they are available for the entire period of interest. For the year 2023, Sant'Agata Bolognese station lacks relative humidity data for the whole period, while Bologna Urbana station has no temperature data. San Pietro

Capofiume station has no data for certain days of June and July 2023. For this reason, in the comparisons in Chapter 4, only the stations for which the data is fully available in the interest period are shown in the figures.

Reanalysis data

ERA5 reanalysis data (C3S 2018; Hersbach et al. 2020) is considered a benchmark when evaluating the performance of the simulations. It has hourly data with a resolution of 31 km, and it is produced with the IFS CY41R2 model, with a modern 4D-var data assimilation scheme. It shows a good representation of low-frequency variability and the temperature anomaly patterns.

I use ERA5 data in the sensitivity tests phase to initialize the nested regional model. This is helpful to test whether the different possible configurations of the downscaling provide any improvement. The results of these tests are discussed in Sections 3.1 and 3.2. Next, an elaboration of ERA5 is also used to complement the seasonal forecast dataset. This is necessary due to the lack of essential surface fields in the seasonal forecast data freely available for download from the Climate Data Store (*CDS*) portal. Specifically, I am introducing from ERA5 the skin temperature, the four volumetric soil moisture levels and three out of the four soil temperature levels. Further details on the dataset preparation are provided in Section 2.2. An overview of the variables available in each dataset is shown in Table 2.1.

Gridded observations data

E-OBS is a 20-member ensemble dataset of daily gridded observations covering most of the European continent and parts of Northern Africa and the Middle East. In this work, I am using version 29.0 with the 0.1 degree regular grid for daily mean temperature. The mean of the ensemble is provided as a best-guess field and is thus used as a reference other than the ground observations when evaluating the simulations. The ensemble spread is also provided. It is estimated by the difference between the 95th and 5th percentile, thus representing a measure of the 90% uncertainty range.

It should be noted that no homogeneity corrections are applied to the station data used to construct E-OBS. Moreover, the number of stations varies over time. For this reason, the use of this dataset is not recommended in evaluating trends.

The E-OBS dataset is developed within the EU-FP6 project UERRA and with the Copernicus Climate Change Service. The data is provided thanks to the ECA&D project. Further details are available in Cornes et al. 2018.

Seasonal forecast data

As a seasonal forecast, the ECMWF system is used. The dataset consists of 25-member hindcasts from 1993 to 2016 and real-time 51-member forecast since 2017. Further details on the model are provided in Section 2.3.

For the first tests, the data is taken from the model run initialized on the first day of June 2023 and considered for three months lead time. The results are discussed in Section 3.2. For the subsequent simulations, which results are explored in Chapter 4, the run initialized in May is used.

At the time of writing, every field is available on a sub-daily basis, every twelve hours on middays and midnights, except for snow depth and sea-ice cover which are instead only daily Copernicus Climate Change Service 2018. All available pressure level data is fed into the WRF model. The missing surface data required are introduced from the ERA5 reanalysis, as anticipated in Section 2.1.

The seasonal forecast data initialized in the months of May and June over the reference period $1993-2016$ is also used to estimate its mean bias during the 2023 summer months and compare it to the one of the ERA5 reanalysis. Further details on the subsequent statistical correction can be found in Section 2.6. The plots representing the mean anomalies are instead shown in Section 3.4.

2.2 Dataset pre-processing

As anticipated in Section 2.1, the seasonal forecast data available on CDS contains fewer surface variable fields compared to ERA5. To perform the downscaling without changing the WRF setup, some missing fields are thus introduced from the latter.

Skin temperature, and the missing soil temperature and volumetric soil moisture levels from ERA5 dataset are remapped using the Earth System Modelling Framework (ESMF) software, to upscale them to the seasonal forecast grid. This new set of fields is then combined with the other existing fields from the seasonal forecast dataset. The missing relative humidity in the vertical levels is computed from the available specific humidity using the MetPy package (May et al. 2022).

Two variations are tested. In the first one, referred to as *option A*, the ERA5 soil data corresponding to the initialization date of the forecast is used to complement the dataset. This introduces a boundary condition that is constant for the initially missing fields in the seasonal forecast dataset. The alternative *option B* consists in building a sub-daily climatology dataset, using ERA5 data from 1993 to 2016. A correction is applied by considering the mean anomaly in the ERA5 fields for the first month of the seasonal forecast. I use Option A for the forecast initialized in June 2023 and Option B for the forecast initialized in May 2023.

The dataset is only partially sub-daily and some fields like snow depth and sea-ice cover are only available daily. The snow depth and sea-ice cover values are assigned to the midnight of each day and linearly interpolated to compute the missing midday value. This is a rough approximation but given the season of interest, I do not expect a significant variation of those variables during the period of interest.

The missing surface pressure is computed using the Hydrostatic Equation 2.1.

$$\frac{dp}{dz} = -\rho g \tag{2.1}$$

The temperature profile is approximated with a constant value, given by the average between the virtual temperature of the first pressure level of the seasonal dataset and the extrapolated one at the surface. The vertical displacement is equal to the difference between the geopotential height of the first pressure level and the model elevation for the corresponding location. The temperature gradient is assumed to be constant and equal to $6\frac{K}{Km}$. The values are thus computed as indicated in Equation 2.2, where T_{1000} indicates the temperature at the lower pressure level of the model. Using this estimate for the temperature it is possible to correct the mean sea level pressure, which is available in the dataset, to obtain the surface pressure, with the expression shown by Equation 2.3, where M_d is the molar mass of dry air and R_d is the specific gas constant for dry air. For a discussion on the lapse rate, the equations in use and their derivation, refer to Wallace and Hobbs 2006.

$$T_{ref} = \frac{T_{1000} + (T_{1000} - \frac{\partial T}{\partial z} \cdot \delta z)}{2} \tag{2.2}$$

$$P_{surf} = P_{msl} \cdot \exp(-\frac{z_{surf} M_d}{R_d T_{ref}}) \tag{2.3}$$

Since these operations alter the standard structure of the dataset, it is no longer possible to solely rely on the WRF preprocessing system (WPS). More specifically, the role played by the *ungrib* component is replaced by an external Python library called *pywinter* (Suárez 2021), which allows building the WRF-WPS intermediate file directly from a dataset in *NETCDF* format.

2.3 Models

SEAS5

SEAS5 is the fifth-generation seasonal forecast system introduced by ECMWF (Johnson et al. 2019). It has a 1°x1° horizontal resolution and uses time steps of 20 seconds. The forecast consists of a 51-member ensemble based on the IFS Cycle 43r1 Model, a spectral model with upper air fields originally output as spherical harmonic coefficients. The ensemble is defined through the perturbation of upper air variables and a small number of land fields, including soil moisture, soil temperature and skin temperature.

The member 0 is the control run, initialized with no perturbations. Further details on the computation of perturbations can be found in ECMWF 2016.

For the hindcasts, the model is initialized with ERA-Interim data, which also drives the 43r1 Surface Model. In forecast applications, the initialization is provided by ECMWF operations.

Manzanas, Torralba, et al. 2022 evaluates the reliability of the SEAS5 forecast system over different areas of the globe and assesses the sensitivity to region definition, hindcast length and ensemble size. Here it is observed that reliability tends to increase with a larger number of ensemble members as they allow a better representation of the uncertainty. Of particular significance for this study is the good reliability observed for forecast temperature during summer over Europe, rendering it an adequate choice as a driving model for the region of interest.

OCEAN5

OCEAN5 is a modern operational ocean analysis system based on the Nucleus for European Modelling of the Ocean (NEMO) version 3.4 model. It provides ocean and sea-ice initial conditions for SEAS5 forecasts. OCEAN5 contains a 5-member ensemble analysis, generated by perturbations to the assimilated observations and to the surface forcing fields. For the ocean ensemble, no unperturbed control forecast is provided.

Each SEAS5 member is assigned to an OCEAN5 member, and then further perturbations are applied, except for member 0. The atmosphere and ocean are coupled hourly to allow the diurnal cycle to be resolved.

WRF

As a regional model, I used the fourth version of WRF, the Weather Research and Forecast model (Skamarock et al. 2019). It has been developed both for research purposes and for numerical weather predictions. The vertical coordinates are based on hydrostatic-pressure and are terrain-following, but they differ from the traditional sigma coordinates, as they remove the influence of the terrain more rapidly while moving to higher altitudes.

The dynamical solver integrates the Euler equations in their compressible, non-hydrostatic flux-form. They include the parametrized physics. The equations being resolved are first rewritten in terms of perturbation variables, to reduce truncation and rounding errors. The newly defined variables are perturbations of a reference hydrostatically balanced state, which satisfies the governing equations for an atmosphere at rest. The equations for the conservation of potential temperature and moisture remain instead unaltered.

The low-frequency modes are integrated using a third-order Runge-Kutta scheme. The higher frequency modes use a smaller time step to maintain numerical stability. A forward-backwards time integration scheme is used for the horizontal propagating acoustic and gravity waves, whereas for the vertical propagating acoustic waves and buoyancy waves, an implicit scheme is preferred. The spatial discretization is instead based on a C grid staggering, meaning normal velocities are staggered half grid with respect to the thermodynamic variables.

WRF is an atmospheric model and without the optional modules for ocean coupling the representation quality of air-sea interaction is limited.

2.4 Simulation setup

To determine the appropriate choice for the WRF configuration a few different tries have been made on a short one-week period. In these preliminary tests, ERA5 reanalysis is downscaled. All the different simulation setups and their characteristics are listed in Table 2.2.

The base simulation, indicated with *Base* and which variations will be discussed afterwards, consists of three nested domains of 150 points per side each, with a grid spacing of nine, three and one kilometre respectively, shown in Figure 2.2. It is centred around the city of Bologna and the larger domain comprises northern Italy and most of the alpine region. It has forty-five vertical levels with a constant surface stretch factor. While it is assumed that the smaller domains can resolve convection in a satisfactory manner, the Kain-Fritsch Parametrization is chosen for the largest one. It is a mass flux parametrization that uses the Lagrangian parcel method to estimate the existence of instability and its availability for cloud growth. More details and the major modifications introduced since its first formulation are available in Kain 2004. As for the radiation option, for all domains and both long-wave and short-wave radiation, the Rapid Radiative Transfer Model (RRTM) is used (Mlawer et al. 1997). The boundary layer scheme is based on the findings from Bougeault and Lacarrere 1989, which extends turbulence parametrization techniques to orography-induced turbulence. The update of the SST is turned off. To improve the representation of the terrain, sixty-one land categories from the USCS dataset are introduced.

The first variation of this configuration is the introduction of coastline interpolation to avoid unreasonable outcomes for the surface variables closer to coastal areas (referred to as *Coast*), like unjustified sharp linear boundaries. This same simulation is then also repeated with an alternative radiation option, namely the RRTMG, which is a more efficient version of the RRTM, thought to have a minimal loss of accuracy for general circulation applications (*Coast_RAD*). In addition, an alternative boundary layer option is explored, the Mellor-Yamada-Janjic (MYJ) scheme (*Coast_PBL*). A further improvement attempt is based on the introduction of a specific urban parametrization approach (*URB*), evaluated for the city of Bologna in Zonato et al. 2020.

Both the basic simulation with only the coastal interpolation and the more sophisticated one with the urban parametrization implemented, are then considered again with a

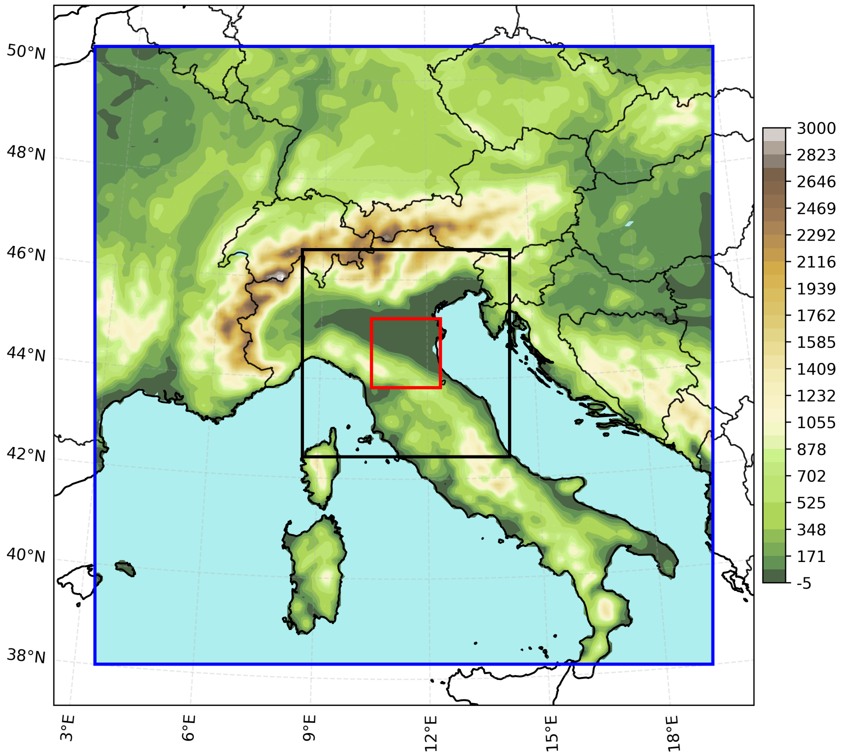

Figure 2.2: Elevation map with Base simulation nested domains. Each domain has 151 grid points per side, distanced by 9, 3 and 1 km respectively.

change to the vertical levels (*Coast_VL* and *URB_VL* respectively), which allows having the first one closer to the ground, at 10 metres, with an increasing distance between the layers as the height increases. This reduces the gap between the point where the temperature is explicitly computed by the model and the one to which the temperature value is extrapolated. This can significantly affect the value, and it is a crucial aspect to test in a comparison with ground station data. Two different configurations are evaluated, with a different number of levels and ground distance of the lowest one (see the alternative with *Coast_VL_Bis*).

Furthermore, a double nesting approach is assessed, with a five kilometres grid spacing for the larger one, and one kilometre for the smaller one. This is not ideal, as it breaks the recommended 1:3 ratio for WRF nestings. However, it is tested in the hope of reducing the computation time while still obtaining meaningful results. The time between radiation physics calls is adjusted accordingly as recommended by the WRF User Documentation. A lower number of vertical levels is used, and the parametrization of convection is turned off. This way there is no intermediate nested domain using a convection scheme.

Seasonal downscaling setup

This time the nestings need to be set up differently, given the coarser grid of the seasonal forecast that is fed to the model. The centre remains the city of Bologna but this time the three nestings have a grid spacing of 27 km, 9 km and 3 km respectively, as shown in Figure 2.3. The configurations are chosen among the ones cited in Section 2.4 according to the verification results shown in Chapter 3.

A variation of the best-performing setup with a continuous update of the SST is also introduced. The impact of this change is assessed in Section 3.4.

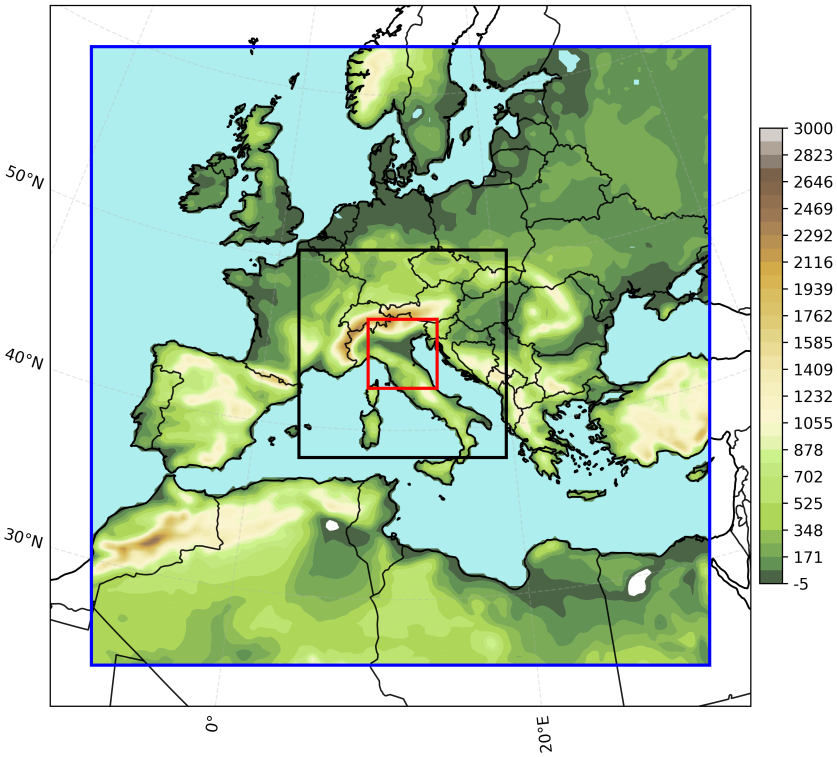

Figure 2.3: Elevation map with seasonal forecast nested domains. Each domain has 151 grid points per side, distanced by 27, 9 and 3 km respectively.

2.5 Expected limitations

The findings in this thesis are exploratory, as they currently lack generality. Only one year and one season are considered. This means the results are not representative of the forecast performance. Moreover, only a handful of members are downscaled, ignoring the others. A visualization of the fraction of data used for this work is shown in Figure 2.4.

A more comprehensive dataset is fundamental for result robustness, since the forecast reliability in a small hindcast may be overestimated, as explained in Manzanas, Torralba, et al. 2022.

Even the testing setup is limited, as it considers only a small set of locations. Moreover, it focuses on the surface temperature, which reliability is usually higher in seasonal forecasts (Manzanas, Torralba, et al. 2022). While this is already helpful in the seasonal prediction of heatwaves, it still tells nothing about precipitation, which correct forecast is just as important.

From the point of view of the model configuration, more initialization options at different lead times should also be explored in subsequent work.

The use of the built-in SST update in the model is still an approximation. The coupling with a true oceanic component in the model would be required to enhance the forecast. This is especially important for those areas where the impact of the SST is more relevant. One further improvement that can be explored in the future is a bias correction of the SST before the data is fed to the downscaling model. The correction of the global model output before its use is a common practice and there are multiple approaches available, as expounded in Appendix B.

The subsampling considers the first five elements of the forecast ensemble, which is an arbitrary choice. A process-informed subsampling has yet to be implemented, but further research is required to deepen the understanding of the local predictors. Considering a broad set of years within a cross-validation approach (Wilks 1995) is vital to improve the choice robustness. It allows a better estimation of the forecast skill and has long been a common practice (Francis and Renwick 1998).

2.6 Benchmark statistical bias corrections

The correction techniques used for the calibration of seasonal forecasts can be classified into different categories: simple bias adjustment methods, ensemble recalibration techniques and more complex statistical downscaling methods. These techniques can fully replace the dynamical downscaling or work in synergy with it to enhance the results. By pre-processing the data that is being fed to the downscaling model it is possible to improve the overall result without a significantly higher cost. They can also be used as post-processing techniques to address known biases in model output.

The simpler techniques correct the mean bias of the GCM and are usually based on a linear scaling, like the one in Lenderink, A. Buishand, and Van Deursen 2007. Further details are provided in Appendix B.

In the present work, I took a reference period spanning from 1993 and 2016. I computed a reference summer two-metre temperature field using seasonal hindcast for that period. Using this average field I computed the anomaly field of the forecast for the year of interest, namely 2023. For each location, the temperature values are then

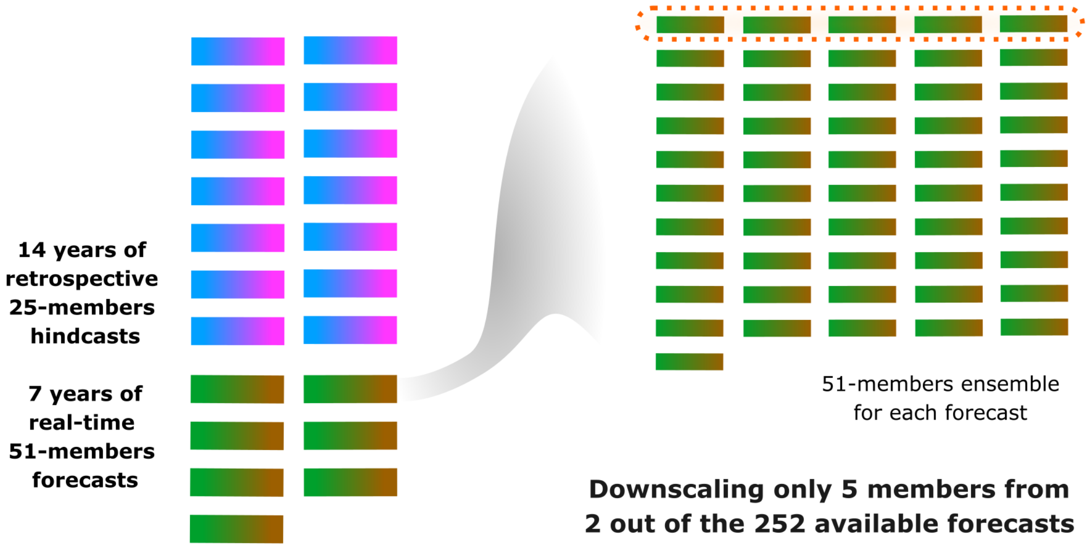
14 years of retrospective 25-members hindcasts
7 years of real-time 51-members forecasts
51-members ensemble for each forecast
Downscaling only 5 members from 2 out of the 252 available forecasts

Figure 2.4: Visualization of the available data for use. The downscaling in this work only considers 5 members of the forecasts initialized in May and June 2023.

adjusted based on the value of the closest grid point of the anomaly dataset, as shown by Equation 2.4. The overline indicates the temporal mean over the three summer months, computed using only midday and midnight values to reduce the computational cost and storage required. The apostrophe marks the corrected field. This scaling does not require any observational reference. It is a rough correction, as it only corrects the mean bias over a certain location.

$$x'_{seas} = x_{seas} + (x_{seas} - \overline{x_{seas}}) \tag{2.4}$$

An alternative option is the Mean and Variance Adjustment (MVA), which uses the data from the observations to correct both mean and variance biases, as in Equation 2.5. Its first application to seasonal forecasts dates back to Leung et al. 1999. The performance of MVA is equivalent or superior to even more sophisticated techniques in multiple instances, as I outlined through a brief review in Appendix B.2. More specifically, in Sections B.2, B.2 and B.3 where more complex methods are explored, their comparison with MVA is simultaneously outlined.

$$x'_m = (x_m - \overline{\langle x_m \rangle})\frac{\sigma_o}{\sigma_m} + \overline{\langle x_o \rangle} \tag{2.5}$$

In my case, I considered a reference period spanning from 2014 to 2022, as the data for each evaluated in-situ station is available for the majority of this period. I use this correction as a surrogate statistical downscaling and compare it to the dynamical downscaling in Chapter4.

More comprehensive bias adjustment methods can still lead to better results if properly calibrated and should be evaluated for future developments. A review of the possibilities is presented in Appendix B. Beyond bias adjustments, more sophisticated techniques are also possible. The two main proper statistical downscaling approaches are the Model Output Statistics (MOS) and the Perfect Prognosis (PP), which I will further detail in Appendix B. Statistical downscaling techniques are essentially empirical relationships between a coarse grid predictor and a local predictand of interest. They could be a cheaper alternative to the dynamical downscaling, but they need calibration. They can be used as a benchmark to evaluate the performance of a dynamical downscaling, as seen in G. T. Diro, Tompkins, and Bi 2012. Another option is blending them with the dynamical downscaling to build hybrid methods. Slater et al. 2023 presents a review of the different ways to achieve this.

2.7 Metrics and indicators

The main scope of this Section is to describe the metrics, indicators and verification modalities in use in Section 3. The verification setup is presented in Section 2.7.1. Moreover, this Section also presents a state of the art of the evaluation metrics that are commonly used in the context of ensemble seasonal forecasting, outlining verification steps to be considered for further evaluation of the proposed methodology. This is done through Sections 2.7.2 and 2.7.3. Applying all of them to the present work would go beyond the intent of this thesis, for which only a set of simpler metrics will be used.

2.7.1 Verification of downscaling

After any adjustment process, it is necessary to evaluate the quality of the forecast, which also means assessing the quality of any downscaling method or bias correction technique applied. As indicators for this first phase, I consider both two-metre relative humidity and temperature. As metrics, I propose the use of two WMO-recommended scores for the verification of a deterministic forecast, namely BIAS and Mean Absolute Error (MAE). In the evaluation of model output against ground observations, the stations should have at least 90% data availability during the verification period.

BIAS and MAE estimate the model accuracy, defined as the distance between the model forecast and actual ground observation. Their expressions are shown by Equations 2.6 and 2.7 respectively, where n is the number of available observations. The BIAS is defined as the mean difference between the forecast values and the observed ones or the reanalysis. The MAE is the average absolute difference between the forecast values and the observations. Often they are computed together with RMSE, Equation 2.8, which is the square root of the mean of the squares of those differences.

In this context, BIAS and MAE are computed to estimate a displacement of the simulation from the in-situ observations and help evaluate its performance. They shall not be considered as an attempt to evaluate the real BIAS and MAE of the model for the area taken into consideration, as only a fraction of the data is considered, as anticipated in Section 2.5. Therefore, the BIAS computed here is not a reliable estimate of the systematic error of the model.

$$BIAS = \frac{1}{n}\sum_{i=1}^{n}(Mod_i - Obs_i) \tag{2.6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Mod_i - Obs_i| \tag{2.7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Mod_i - Obs_i)^2} \tag{2.8}$$

Correlation is a measure of association, which shows the strength of linear relationships between forecast and observation. It can also be computed considering the respective anomalies. It is shown in Equation 2.9, where the overline indicates the sample mean.

$$r_{Pearson} == \frac{\sum_{i=1}^{n}(Mod_i - \overline{Mod})(Obs_i - \overline{Obs})}{\sqrt{\sum_{i=1}^{n}(Mod_i - \overline{Mod})^2 \sum_{i=1}^{n}(Obs_i - \overline{Obs})^2}} \tag{2.9}$$

The first tests are performed by initializing the model with ERA5 reanalysis data for the 21st of August at $00:00$. The model is set to run until the 28th of August at $12:00$ for a maximum of 24 hours, within which nearly all the simulations are completed. This is done to assess the ability of the simulation to correctly represent a heatwave scenario, and the fourth week of August 2023 has temperatures well above the climatological average for the period over northern Italy, which is the area of focus. BIAS, MAE and Pearson Correlation are computed using instantaneous hourly values, as it is the minimum common frequency provided for the temperature and relative humidity data across the set of weather stations. The same scores are also computed for the daily extremes. The most relevant results, alongside the time series, are presented in Section 3.1.

A further comparison follows in Section 3.2, considering the entire three months of June, July and August 2023. This evaluates the ability of the downscaling setup to represent surface temperature across the entire season. It verifies if what is observed for the weekly comparison is confirmed for this larger time range, which includes a more diverse set of meteorological conditions. The seasonal forecast downscaling and the impact of a continuous SST update are evaluated in the same way on one member in Sections 3.3 and 3.4. A wider overview with multiple members, using BIAS and MAE of monthly-averaged two-metre temperature, is instead shown in Chapter 4.

As suggested in *Manual on the WMO Integrated Processing and Prediction System* 2023, these computations shall happen both aggregating the whole domain and for every individual grid point and this helps assess the spatial variability of the forecast skill. For the present study, I only consider a set of locations, among the ones described in Section 2.1. While focusing only on the area around the city of Bologna, this approach still allows the evaluation of potential differences between rural and urban cases.

A comprehensive verification of seasonal forecasts would encompass a comparison of the spatial patterns across the ensemble (G. T. Diro, Tompkins, and Bi 2012), with specific attention to the extremes. This is especially important for those regions with complex topography in which we would expect to see greater benefits from the introduction of a dynamical downscaling method.

2.7.2 Verification of the subsampling technique

At the time of writing, there are no universal verification strategies for the evaluation of subsampled seasonal forecasts. Ensemble mean and spread are key figures to consider, and the latter can be estimated using the interquartile range, as in Park and Kam 2023. With the ensemble mean it is possible to compute the scores presented in Section 2.7.1.

A simple Mean Square Error Skill Score (MSESS) in combination with the Heidke Skill Score (HSS) have been recently used by Kowal et al. 2024. Together they are considered a suitable choice, since the former allows the investigation of the overall quality of the forecast, while the latter measures the ability to predict extreme events, once a quantile threshold has been defined. These simple metrics can be used to easily assess different subsampling strategies before the final choice is made. Equation 2.10 shows the HSS expression considering the four coefficients of a standard contingency table for binary events. Furthermore, the MSESS together with its three-term decomposition, are a common choice in the verification of deterministic subseasonal and longer-range forecasts, as recommended in *Manual on the WMO Integrated Processing and Prediction System* 2023. More details on the decomposition of scores are presented in Appendix A. The MSESS is obtained by subtracting to 1 the ratio between the mean square error of the forecast and that of the reference, which could be the climatology. The formula is shown as Equation 2.11. If we assume the model to be perfect, it is also possible to compute the ensemble mean MSSS starting from the individual members, as explained in J. M. Murphy 1988. Then, it is possible to express it in terms of the mean MSESS of the ensemble members, and it is ensemble size-dependent. This can be helpful to effectively estimate the skill of the forecast system, as shown for the troposphere in J.-Y. Han et al. 2023.

$$HSS = 2\frac{ad - bc}{(a+c)(c+d) + (a+b)(b+d)} \tag{2.10}$$

$$MSESS = 1 - \frac{MSE_{forecast}}{MSE_{reference}} \tag{2.11}$$

Another option is to compare the estimated observation spread with the error associated with the forecast, computed in F. Pappenberger et al. 2009 as a simple difference with the observations. In that study, the spread difference was plotted against the mean difference between the observations and the forecast. A predominance of positive, or negative, differences in the spreads indicates an overestimation, or underestimation, of the spread by the forecast.

2.7.3 Probabilistic metrics

To further take advantage of the ensemble approach, it is possible to compute scores which consider the members in a probabilistic framework, as suggested by G. T. Diro, Tompkins, and Bi 2012. This allows considering the distributions of the forecast variables across the entire ensemble. These include the Rank Probability Skill Score (RPSS) and the Relative Operating Characteristic Skill Score (ROCSS). They are commonly used in seasonal forecast evaluation (Weisheimer and Palmer 2014; Manzanas, Lucero, et al. 2018; Manzanas, J. Gutiérrez, et al. 2018), together with reliability diagrams.

Measuring accuracy

The Relative Operating Characteristic (ROC) measures the ability of the forecast to correctly discriminate different categorical events (Mason 1982). Events are commonly classified using terciles (Manzanas, J. M. Gutiérrez, Bhend, Hemri, F. J. Doblas-Reyes, Torralba, et al. 2019). The ROC curve represents the hit rate corresponding to a certain false alarm rate, providing information also regarding the extremes of the variable distributions. The area underneath this curve (AUC) can be used to decide whether the forecast can be considered skilful. AUC should be higher than 0.5, and the closer it is to one, the better the quality of the forecast.

The Relative Operating Characteristic Skill Score (ROCSS) is based on the concept of ROC and quantifies the improvement with respect to a random classification of the events. It has been used as a measure of accuracy in the context of seasonal temperature forecasts (Manzanas, J. Gutiérrez, et al. 2018) and is regarded as a reasonable index to communicate the value of a forecast (Manzanas, Frías, et al. 2014), other than being a recommended verification system for long-range forecasts by the bom.gov.au. Using a ROC diagram with standardized AUC is also recommended by *Manual on the WMO Integrated Processing and Prediction System* 2023 for the verification of subseasonal and longer-range forecasts.

Measuring reliability

Reliability is a measure of how close the observed frequency of a certain occurrence is to the forecast probability, and it is built using the different members of the ensemble forecast system. The Rank Probability Score (RPS) (Epstein 1969, Wilks 1995) can be used to measure the agreement of the forecast with the observations.

This score requires the definition of categories in which the observations fall, like in Kumar, Barnston, and Hoerling 2001 and Doblas-Reyes et al. 2009. For instance, a common choice is the introduction of terciles or quartiles. For each of the ranges, a probability is computed from the ratio between the number of ensemble members falling within it over the total. This is a way to quantify the probability of each class of events, which is then used to assess its accuracy, namely how often the observation matches with the most likely event according to the forecast. The RPS represents the deviation of the forecast from what would have been the correct classification of the occurrences. Its value ranges from zero for a perfect forecast, in which every value falling within a certain bin corresponds to an observation in that quantile range, to one for an entirely wrong one.

Equation 2.12 represents the RPS, where f_j are the forecast probabilities while e_j represent the reference ones. The parameter c stands for the number of categorical events. Using the RPS it is possible to define the Rank Probability Skill Score (RPSS). Its value ranges from one for a perfect forecast to minus one for a completely misleading forecast, where a value equal to zero stands for a result equivalent to considering the local climatology. To de-bias the reference RPS when computed using observations of the local climatology, Tippett 2008 presented a variation of the RPSS definition, which provides an unbiased score in the limit of an infinite number of ensemble members. Equation 2.13 shows the corrected expression.

$$RPS = \frac{1}{c-1} \sum_{i=1}^{c} (\sum_{j=1}^{i} f_j - \sum_{j=1}^{i} r_j)^2 \tag{2.12}$$

$$RPSS = 1 - \frac{RPS_{forecast}}{RPS_{reference} + \frac{1}{n_{members}} RPS_{reference}} \tag{2.13}$$

There exists also a continuous version of this skill score, namely the Continuous Rank Probability Skill Score (CRPSS). It can be computed from the Continuous Rank Probability Score (CPRS). This generalized version does not require the definition of quantile categories. In the context of global ensemble forecasts, *Manual on the WMO Integrated Processing and Prediction System* 2023 recommends the computation of the CPRS for both the ensemble prediction system and the control deterministic forecast, in which case it is equivalent to the mean absolute error (MAE).

The particular case of the RPS for binary categorical events is the Brier Score (BS). The Brier Skill Score (BSS) has long been used as a measure of reliability (Brier 1950), and its computation is required by *Manual on the WMO Integrated Processing and Prediction System* 2023 for every variable in the verification of the global ensemble prediction system. However, even a perfectly reliable forecast is useless if it always provides the same probabilities as the climatology. Moreover, a forecast with a low score might still be useful.

To avoid erroneously discarding them, Weisheimer and Palmer 2014 proposed a five-categories system. Firstly, reliability is evaluated for multiple classes of events, using the forecast probability of a certain occurrence and its expected long-term climatological frequency. The types of events are usually defined using a tercile threshold approach, which is also useful as it eliminates the influence of BIAS (Weisheimer and Palmer 2014). Each of them is then plotted on a reliability diagram with a frequency histogram. Then, to evaluate the overall reliability of the forecast, a linear regression is introduced, and the slope becomes an indicator for reliability, with a perfect forecast having a value of one for it. As the slope approaches zero, the added value of the forecast diminishes compared to a reference climatological forecast. If the value of the slope is negative, the forecast is said to be worse than useless, as it shows an opposite correlation between the forecast probability of an event to happen and its climatological frequency.

The value is associated with uncertainty which can be estimated through a bootstrap algorithm (Weisheimer and Palmer 2014). The categories are defined based on the estimate for the slope and its range, for which a 75% confidence was proposed by Weisheimer and Palmer 2014, but more recently a more conservative 90% has been used by Manzanas, J. M. Gutiérrez, Bhend, Hemri, F. J. Doblas-Reyes, Torralba, et al. 2019. The best possible scenario is a range containing a slope equal to one, associated with perfect reliability, and fully contained within the region contributing positively to the computation of the BSS. The forecast system is said to be still useful when the slope is higher than 0.5, but its range does not include the perfect reliability line. A marginally useful forecast is one still having a significantly positive slope, while none of the previous conditions is met. This class was further split into two (Manzanas, Lucero, et al. 2018) based on the belonging or not of the uncertainty range to the positive skill region. A non-useful forecast system has a slope compatible with zero while for a negative slope, the forecast is said to be dangerously useless. It is crucial to have a long hindcast to provide robust estimates for reliability, as shown by Manzanas, Torralba, et al. 2022.

Dataset	**Seasonal forecast sub-daily data**	**ERA5 hourly data**
10m u component of wind	A	A
10m v component of wind	A	A
2m dewpoint temperature	A	A
2m temperature	A	A
land-sea mask	A	A
mean sea level pressure	A	A
sea ice cover	A	A
sea surface temperature	A	A
skin temperature	M	A
snow depth	A	A
soil temperature level 1	A	A
soil temperature level 2	M	A
soil temperature level 3	M	A
soil temperature level 4	M	A
surface pressure	M	A
volumetric soil water layer 1	M	A
volumetric soil water layer 2	M	A
volumetric soil water layer 3	M	A
volumetric soil water layer 4	M	A
geopotential	A	A
relative humidity	M	A
specific humidity	A	A
temperature	A	A
u component of wind	A	A
v component of wind	A	A
pressure levels	12	37

Table 2.1: Variables required for the WRF model initialization and their availability in the different datasets. *A* indicates availability of the field, while *M* indicates the data is missing from the seasonal forecast dataset. The number of pressure levels is also shown at the bottom of the table for each dataset.

Name	**Number of x-y grid points**	**Grid spacing (km)**	**Features and differences with Base configuration**
Base	150 - 151 - 151	9 - 3 - 1	61 land use 45 vertical levels
Direct	600	2	Time step of 6 seconds
Coast	150 - 151 - 151	9 - 3 - 1	Coastline interpolation
URB	150 - 151 - 151	9 - 3 - 1	Coastline interpolation Zonato 2020 urban param
Direct_Bis	900	1	Timestep of 1 second
URB_larger	252 - 253 - 253	9 - 3 - 1	Same as URB
URB_5-1	150 - 251	5 - 1	radt = 5 cu_physics = 0 (even at 5 km) 38 vertical levels (lower at 5 m)
Coast_VL	150 - 151 - 151	9 - 3 - 1	53 vertical levels (lower at 5 m)
URB_VL	150 - 151 - 151	9 - 3 - 1	53 vertical levels (lower at 5 m)
Coast_VL_Bis	150 - 151 - 151	9 - 3 - 1	40 vertical levels (lower at 10 m)

Coast_PBL	150 - 151 - 151	9 - 3 - 1	Changed PBL physics
			boundary = 2
			sf_sfclay_physics = 2

Coast_RAD	150 - 151 - 151	9 - 3 - 1	Changed radiation physics
			ra_lw_physics = 4
			ra_sw_physics = 4

Table 2.2: Simulations setup and changed WRF settings from Base setup.

Chapter 3

Verification

The following tests evaluate the ability of the different WRF simulations to accurately reproduce temporal patterns. This assessment is conducted at selected locations within the area of interest where in-situ observations are available.

Firstly, some preliminary sensitivity tests are presented in Section 3.1 to quickly evaluate the differences between different configurations of the WRF model when downscaling ERA5 data for a single week. Then a selected set of configurations is also tested on the entire season in Section 3.2, to check whether the results can be generalized over the longer three-month period. In Section 3.3, the final choice for the WRF configuration is tested for the downscaling of one member of the seasonal ensemble forecast. The impact of updating the sea surface temperature is evaluated in Section 3.4. The scores for ERA5 are also computed, as it is used as a benchmark.

To provide an additional context before continuing, the distribution of summer hourly temperatures for each station are computed. This allows a comparison of the mean ERA5 summer temperature values for each hour of the day with the values provided by each station in the 10 years between 2014 and 2023. The behaviour is location-specific during the night. As an example, the data for Bologna Idrografico (BOI) is shown in Figure 3.1, for which the night temperature is underestimated on average. During the day the temperature provided by ERA5 is smaller than the observed mean of the distributions, except for late afternoon where the ERA5 mean tends to overestimate the mean temperature. In all locations outside the city of Bologna, there is an overall overestimation of temperature, especially evident during the night in San Pietro Capofiume (SPT), shown in Figure 3.2, which is the one located further from the bigger settlements.

3.1 Preliminary sensitivity tests

The simulations are referred to with the acronyms introduced in Section 2.4 and summarized in Table 2.2.

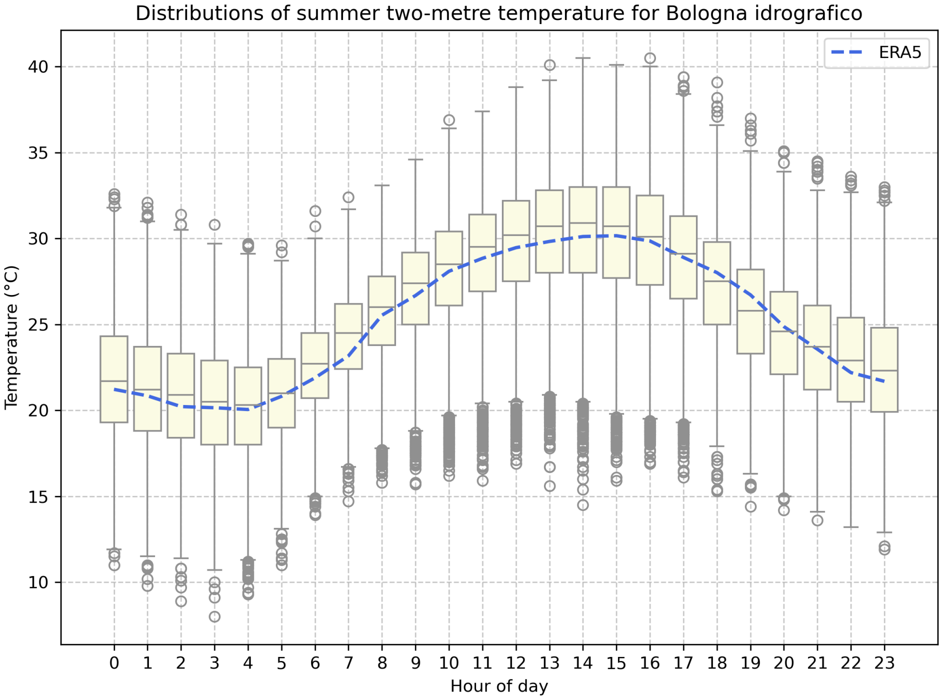

Figure 3.1: Hourly distributions of two-metre temperature as measured by the station located in Bologna Idrografico (BOI) in the period $2014 - 2023$. The dashed line represents the mean value from the ERA5 reanalysis. ERA5 tends to underestimate the temperatures as the dashed line lies below the observed temperature distribution for the majority of the time.

The computational times are similar for all setups, except for the ones with the alternative configuration of the vertical levels, namely *Coast_VL* and *URB_VL*, which take marginally less time to be completed. Other notable mentions are *Direct* and *Direct_Bis*, both of which could not be completed despite the lowered integration time step, and *URB_Larger* which could not be completed within the time constraint given.

To assess the quality of the different simulations, the data from five of the weather stations mentioned in Section 2.1 is considered. From a first look at the time series, it is possible to notice better agreement with ground observations for the two-metre temperature series than for the two-metre relative humidity.

Focusing on the available urban station of Bologna Idrografico (BOI), the temperatures tend to be overestimated during the first two days, especially by *URB_VL*. However, from day three, by which the heatwave intensifies, it is the one that comes closer the observed temperature around the middle of the day. This aspect will be also evaluated more quantitatively in Section 3.1. The mentioned model can not be considered the best overall, since it is also the one which seems further from observations in the second half

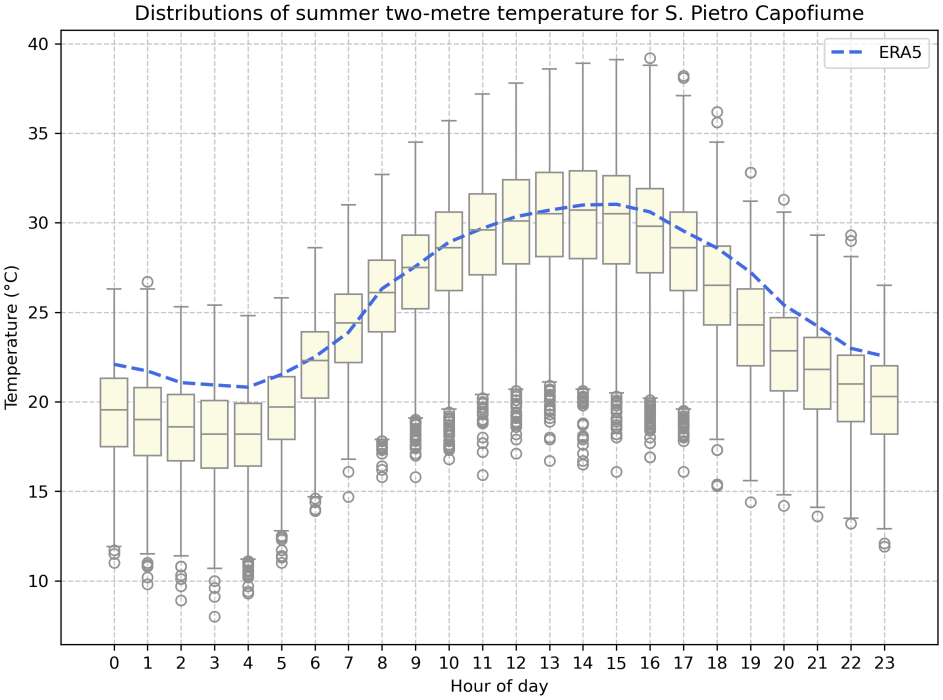

Figure 3.2: Hourly distributions of two-metre temperature as measured by the station located in San Pietro Capofiume (SPT) in the period 2014 – 2023. The dashed line represents the mean value from the ERA5 dataset. ERA5 tends to overestimate the temperatures as the dashed line lies above the observed temperature distribution, especially during the night.

of the day, as shown in Figure 3.3.

The night minima are mostly overestimated, in particular during the central days of the heatwave. On the contrary, ERA5 reanalysis tends to underestimate the temperatures during the night, as confirmed by the overall negative BIAS for this station shown in Table 3.1. ERA5 also shares the same difficulties of the tested simulations in representing the highest daily values. Other features that can be seen in the short period shown should not be intended as a general behaviour of the simulations.

Moving to rural stations, all simulations tend to overestimate the night temperatures. ERA5 behaves similarly. As an example, part of the San Pietro Capofiume (SPT) time series is shown in figure 3.4. The *Coast_RAD* setup, which has the alternative radiation parametrization scheme, and the ERA5 reanalysis are the ones overestimating the most evening and night temperatures. The behaviour is less consistent across stations and different days for the daily maximum temperatures.

As for the surface temperature, even with the two-metre relative humidity the differences between the simulations are more evident during the night. All simulations tend

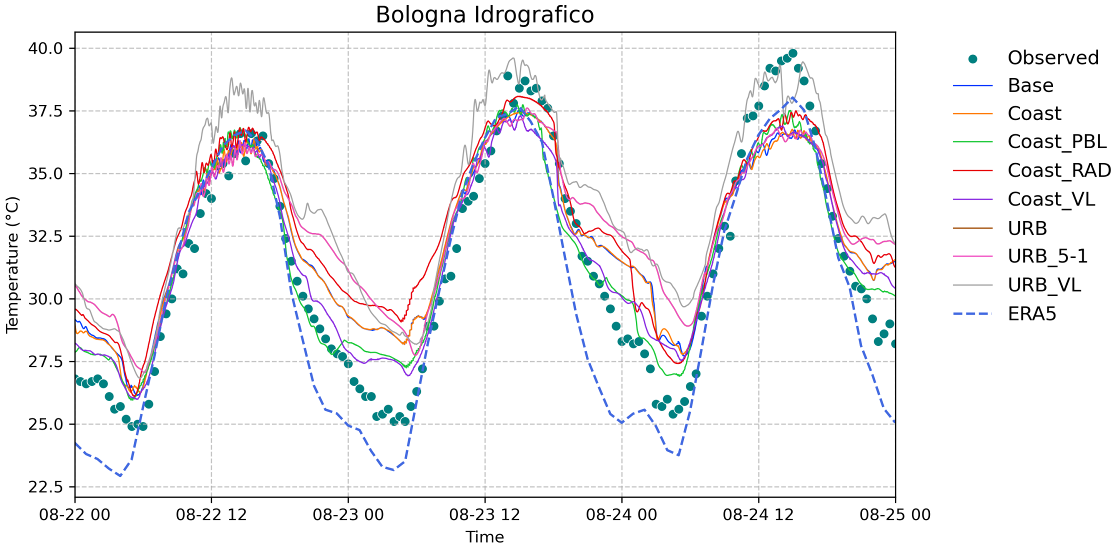

Figure 3.3: Two-metre temperature time series for the Bologna Idrografico (BOI) station between the 22nd and the 25th of August 2023. The different simulations are represented by continuous lines of different colours, ERA5 reanalysis data is shown as a dashed line and the observations are plotted as blue dots.

to underestimate it to a certain degree.

The vertical profiles

A correct representation of the vertical temperature profile is also important to evaluate the correct functioning of the simulations. The vertical profiles for August 22nd at 12:00 GMT and the one for August 23rd at 00:00 GMT are plotted against the observation of the available soundings of the station located in San Pietro Capofiume. The simulations, shown in Figures 3.5 and 3.6, largely follow the observed profile and correctly represent the nocturnal surface inversion, as shown in Figure 3.6. The added value of downscaling the reanalysis is clear close to the ground.

Quantitative outlook

To get a more quantitative outlook BIAS, Mean Absolute Error (MAE) and Pearson correlation are computed using hourly data. More details on the computation are available in Section 2.7.1. The station acronyms follow the convention introduced in Section 2.1.

Let us first consider two-metre temperature BIAS, presented in Table 3.1. It shows the values for each simulation and each of their nestings, indicated with *d01*, *d02* and *d03*, where the latter is the smaller.

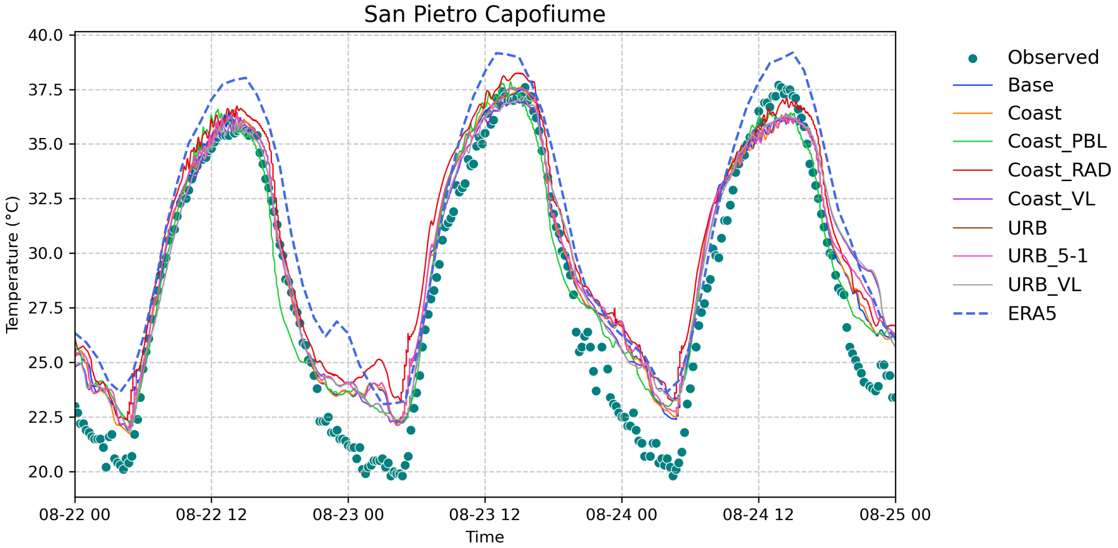

Figure 3.4: Two-metre temperature time series for the San Pietro Capofiume (SPT) station between the 22nd and the 25th of August 2023. The appearance of the elements in this figure matches that of Figure 3.3.

It is clear the advantage provided by downscaling the ERA5 dataset. The BIAS is reduced in all cases, especially in the urban stations. The introduction of nestings with smaller grid-spacing does not consistently improve the score. Among the alternative configurations to the base one, none of them improves the score significantly and coherently across the examined stations. No significant difference is noticed across the board.

Coast_PBL, which has the alternative PBL scheme, is the worst-performing one in the urban context in the *d01* 9 km grid-spacing nesting while providing the smallest BIAS for the rural stations. It then becomes the best one considering the smallest *d03* nesting. The *Coast_VL* and *URB_VL* setups, having the first level closer to the ground, seem to provide a small advantage.

URB_5-1 which has only two nestings is the worst performing of the group. This is no surprise, as it ignores the recommended three-to-one ratio for grid spacing between successive nestings in WRF.

Moving to the two-metre relative humidity BIAS, the conclusions that can be drawn are similar, as it is shown in Table C.2 in Appendix C. In fact, while it is still beneficial to perform the first downscaling, no further improvement is seen with the other two nestings, which in most cases actually worsen the score. The introduction of the specific urban parametrization shown almost no effect on the score, while the effect of the alternative PBL scheme mirrors the one it has on the temperature BIAS.

Next, let us focus on the Mean Absolute Error, which must be considered in combination with BIAS. A reduced BIAS can result from the offsetting of positive and negative

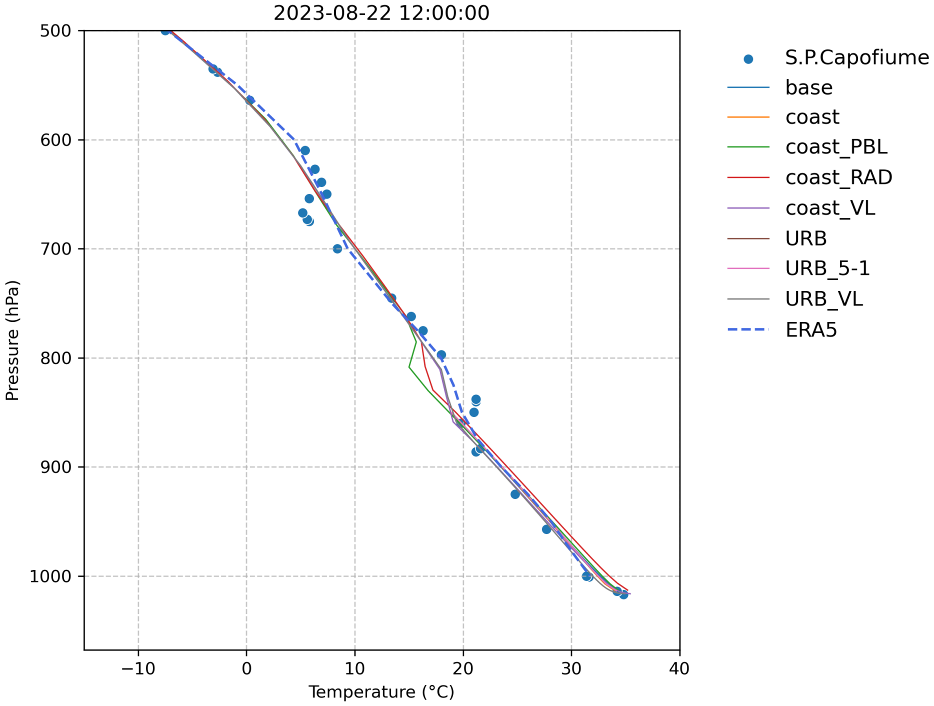

Figure 3.5: Vertical temperature profile for the San Pietro Capofiume (SPT) station at 12:00 GMT on 22nd August 2023. Simulations are represented by continuous lines, the observed profile is shown as dots, and the ERA5 reanalysis profile is shown as a dashed line.

contributions rather than indicating a simulation that is closer to the observations in absolute terms. Just as for the BIAS, there is no clear difference in this score for both variables with the different grid spacings.

When considering the two-metre temperature, MAE is always improved by all downscaling configurations, with the notable exception of *Coast_RAD*, which worsens the score in three out of the four stations analysed, and consistently remains the worst-performing configuration even in the smaller grid-spacing cases. A small advantage is observed with *Coast_VL*, *URB_VL* and *Coast_PBL*. When turning to two-metre relative humidity, a slight reduction is observed for the urban station, while the scores are worse for almost every configuration in the rural ones. The MAE values for temperature and relative humidity, shown in Tables C.3 and C.1 respectively, are available in Appendix C.

As for correlation, the values are already high for ERA5, with values higher than 0.85 for relative humidity and above 0.9 for temperature. After the downscaling the correlation remains high even though it tends to be slightly reduced, especially in the urban context. The alternative radiative transfer scheme tends to be the worst performing and further deteriorates the correlation, with the finer nesting showing a coefficient value

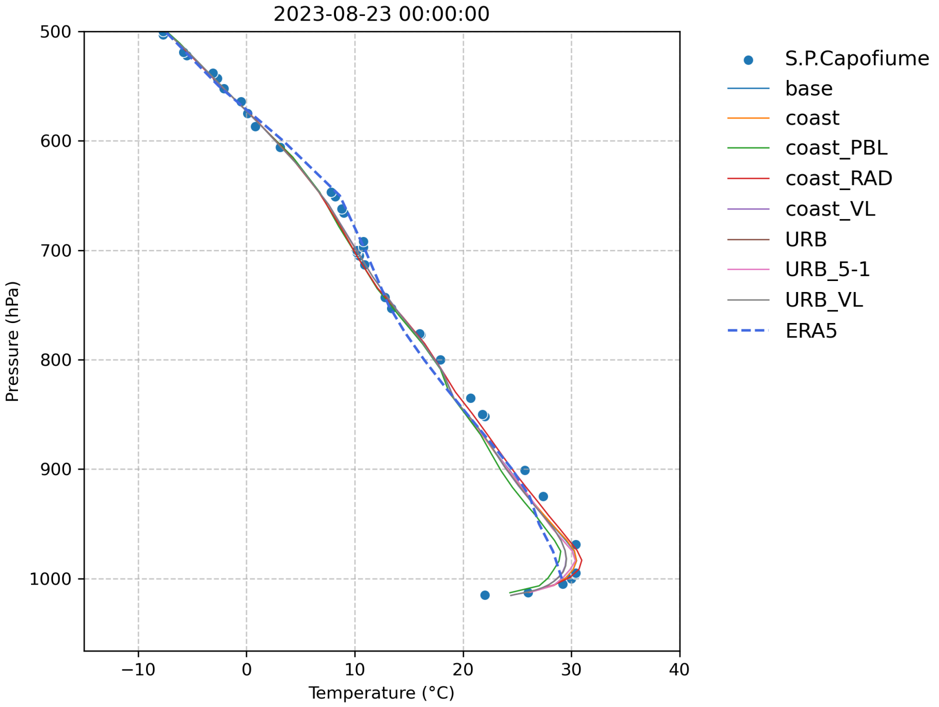

Figure 3.6: Vertical temperature profile for the San Pietro Capofiume (SPT) station at 00:00 GMT on 23rd August 2023. The appearance of the elements in this figure matches that of Figure 3.5.

below 0.7 for the relative humidity. The difference is less clear for the temperature, for which the grid spacing and the configurations chosen seem to have little to no impact on the correlation.

The double-nesting approach, which has not been under the spotlight until now in this section, often performs worse than the others. At the same time, it does not significantly contribute to reducing the computational cost, which was the reason behind its introduction in the first place.

The configuration *Coast_VL* with the first level shifted at 5 metres from the ground increases the number of levels to 53 compared to the *Base* configuration, which has 45. In the alternative configuration of the vertical levels used in *Coast_VL_Bis*, the first level is forced to be at 10 metres, while the total number of levels is set to 40. No significant change is observed by changing the number of levels, as shown with simple scores in the Tables in Appendix C. The observed computational time has not decreased substantially.

To assess whether the boundaries are too close to the region of interest, a larger domain is tested with *URB_larger* to evaluate if it could significantly improve the results. No clear advantage is given by the larger domain. This alternative configuration is also

Simulation	Domain	BOI	MEZ	STG	SPT
Base	d01	0.05	1.14	0.68	1.39
Coast	d01	0.05	1.18	0.69	1.43
PBL	d01	-0.42	0.64	0.20	0.96
RAD	d01	0.63	1.68	1.37	1.97
VL	d01	-0.41	1.01	0.28	1.31
URB	d01	0.06	1.19	0.70	1.40
URB_5-1	d01	1.97	1.57	0.93	1.72
URB_VL	d01	0.27	1.00	0.30	1.28
ERA5	d01	-1.31	1.48	1.62	2.46
Base	d02	0.28	1.45	1.22	1.60
Coast	d02	0.26	1.46	1.20	1.61
PBL	d02	-0.33	0.79	0.45	1.07
RAD	d02	0.86	2.02	1.78	2.18
VL	d02	-0.43	1.04	0.72	1.33
URB	d02	0.47	1.40	1.23	1.57
URB_5-1	d02	2.50	1.78	1.07	1.95
URB_VL	d02	0.92	1.05	0.76	1.37
Base	d03	0.73	1.46	0.90	1.49
Coast	d03	0.72	1.47	0.91	1.51
PBL	d03	0.11	0.85	0.28	1.09
RAD	d03	1.26	2.09	1.49	2.14
VL	d03	0.15	1.10	0.58	1.33
URB	d03	1.02	1.44	0.92	1.53
URB_VL	d03	1.69	1.12	0.62	1.37

Table 3.1: BIAS of two-metre temperature for the different simulations and domains, referred to the selected locations.

associated with a higher computational cost and is thus discarded. The Tables are shown in Appendix C.

A key aspect in the representation of temperature is a correct portrayal of the extrema. Starting with the results for the maximum temperature BIAS, the benefit is unclear. In some cases, it is significantly reduced and brought very close to zero. However, this does not happen consistently across all available locations for any of the presented configurations. Interestingly, the worst performing across the board seem to be *Coast_VL* and *URB_VL*, despite their first level closer to the ground.

The same analysis is repeated with temperature minima. As for the maxima, there is not a consistently better configuration of the downscaling and the further nestings have little to no impact. Most setups can marginally reduce the BIAS, especially *Coast_PBL*. The worst performing one is instead *Coast_RAD*, coherently with the observed time series in Figure 3.4.

To summarize, the downscaling is effective in terms of BIAS reduction. The downscaling has also proven effective in a few cases in improving the scores over the ERA5 value, especially in the available urban station. This aspect is investigated with the addition of another station in Chapter 4. The enhancement is likely due to the higher resolution of the topography and the land use information introduced with the WRF downscaling.

Another conclusion which is possible to draw is the lack of noticeable improvements in the observed scores given by the additional nestings beyond the *d01*. Being the reference urban stations located within the city centre, the advantage of the better land use representation can be already evident from the *d01* nesting, with mostly minor differences with the successive ones. This aspect is certainly setup-dependent, as for example *Coast_PBL* changes significantly its relative performance within the tested batch depending on the nesting considered.

According to Meehl et al. 2021 the increased resolution must also be accompanied by comparable increases in the quality of the physical parametrizations such as cloud feedback and cloud-aerosol interactions. This may be part of the reason why no further improvement is observed thanks to the finer grid spacing of the additional nestings.

While these tests are useful for sounding the possible effect of changes in the setup, no single configuration of the downscaling consistently outperforms the others across all metrics. On the other hand, these sensitivity tests still allow the exclusion of some configurations, such as *Coast_RAD* with the alternative radiative transfer scheme, as it generally performs worse.

3.2 Tests on the entire season

The same triple nesting is maintained in this second comparison too, despite no evidence of an improvement in the scores coming from the use of a finer grid alone. This is done due to the longer time range considered, which consequently comprises a greater variety of situations. Among the previously considered locations, three have temperature data available for the entirety of the period and only for those the scores are computed.

URB which introduces the specific urban parametrization is tested again, given the importance of such variation in the representation of the urban context. To conclude this specific simulation, it was necessary to lower the time step for the last week of August from the 54 seconds used in all other tests to 9. This is due to a higher numerical instability introduced by the urban parametrization. This increased the time required to complete the simulation. *Coast_VL* is also run for this entire period, given its lower observed computational time. This setup required a lowering of the time step as well for the last week. The dynamical time step is reduced to 9 seconds while the one for the radiative transfer is lowered to 3 seconds instead of the 9 seconds of *Base*. These variations caused a significant increase of the computational time.

Let us start by considering the same scores computed in Section 3.1. The relative humidity BIAS is reduced significantly for the urban stations, which initial ERA5 BIAS is much higher than the one seen in the available rural station located in Mezzolara (MEZ). For this station the BIAS is increased by the downscaling. No benefit is provided by additional nestings, as the BIAS generally increases with respect to the simulation with 9 km grid spacing, as shown in Table 3.2.

Simulation	**Domain**	**BOU**	**BOI**	**MEZ**
Coast	d01	-5.56	-7.16	-4.09
URB	d01	-3.23	-4.84	-4.15
VL	d01	-3.12	-4.70	-0.15
ERA5	d01	13.74	12.07	1.39
Coast	d02	-8.83	-10.45	-9.92
URB	d02	-8.45	-10.06	-10.14
VL	d02	-6.71	-8.33	-7.64
Coast	d03	-9.67	-11.21	-10.14
URB	d03	-9.34	-11.07	-10.27
VL	d03	-7.75	-9.34	-7.78

Table 3.2: BIAS of two-metre relative humidity for the different simulations and domains, referred to the selected locations.

Moving to the temperature BIAS, shown in Table 3.3, a reduction is observed in all instances. The additional nestings marginally worsen the BIAS compared to the first one, but the values are still lower than the ERA5 values.

The improvements are much less evident for relative humidity MAE, which is only marginally reduced in the urban stations. Just as for the BIAS, MAE increases for the Mezzolara rural station. Regarding the temperature, MAE is only marginally reduced for the Bologna Idrografico (BOI) station, while there is no clear effect of the application of the downscaling on the other two available locations. The relative humidity correlations are slightly reduced by the downscaling at all stations, from values above 0.8 for ERA5. Small or no changes are seen with temperature correlation, which remains above 0.9. No significant change is observed with finer grid spacing.

Simulation	Domain	BOI	MEZ	STG
Coast	d01	0.37	0.15	0.04
URB	d01	-0.12	0.15	0.05
VL	d01	-0.07	-0.31	-0.45
ERA5	d01	-1.63	0.71	0.85
Coast	d02	0.83	0.92	0.81
URB	d02	0.68	0.93	0.75
VL	d02	0.34	0.64	0.59
Coast	d03	1.21	0.93	0.80
URB	d03	1.09	0.89	0.73
VL	d03	0.82	0.67	0.65

Table 3.3: BIAS of two-metre temperature for the different simulations and domains, referred to the selected locations.

Time series

The limited changes in the scores considered are confirmed by a review of the time series. I compute a daily value for the surface temperature for the downscaled forecast. This statistic is obtained by averaging the midnight and noon values for every location. This simplifies the visualization as opposed to showing hourly data and lowers the computation cost compared to using the whole dataset, as the scope is only to provide a quick way of comparing the temporal behaviour of the different simulations in the context of the reference datasets and in-situ observations. As a representative example, the daily averages for July are shown in Figure 3.7 for the station of Bologna Idrografico. The different simulations are barely distinguishable and largely overlap for most of the days. No single configuration is consistently better than the others.

Extremes comparison

No consistent BIAS or MAE reduction is observed neither for temperature maxima. In most cases, they are instead slightly increased. No significant difference is visible when changing the resolution. The biggest benefit can be observed in the available urban location, but only starting from the *d02* nesting. Maximum temperature BIAS is shown in Table 3.4 for all available locations. The daily maximum temperature correlation are high for ERA5, and remain above 0.9 after the downscaling to 9 km. No effect or a minor worsening of the correlation coefficient is observed with the finer nestings.

Moving to daily minimum temperature, a reduction of BIAS is instead evident across all stations. This is also reflected into a reduction in MAE, not shown. Coherently with

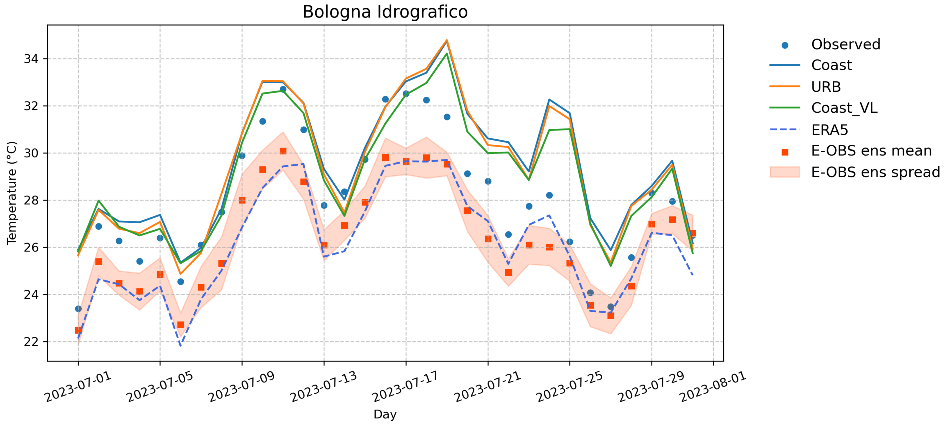

Figure 3.7: Daily two-metre temperature values for the month of July 2023 at the Bologna Idrografico station. The continuous lines indicate the different downscaling simulations, and the dashed line represents ERA5 reanalysis. The blue dots refer to in-situ observations while the orange squares represent the E-OBS dataset. The thicker line represents the non downscaled seasonal forecast.

Simulation	Domain	BOI	MEZ	STG
Coast	d01	-1.51	-0.24	-1.02
URB	d01	-1.95	-0.30	-0.99
VL	d01	-1.99	-1.02	-1.75
ERA5	d01	-1.83	0.17	-0.61
Coast	d02	-0.49	1.34	0.58
URB	d02	-0.89	1.37	0.89
VL	d02	-0.87	0.92	0.16
Coast	d03	-0.01	1.80	0.74
URB	d03	-0.36	1.70	0.94
VL	d03	-0.40	1.05	0.30

Table 3.4: BIAS of daily two-metre maximum temperature for the different simulations and domains, referred to the selected locations.

what is already seen in Section 3.1, no significant improvement is observed with the introduction of the additional nestings. The daily minimum temperature BIAS is shown in Table 3.5. The correlation, not shown, remains moderate to high for all stations, with little impact from the introduction of the downscaling.

The results over the tested three months period show no consistently different behaviour between the different configurations tested. As seen in Section 3.1, no significant

Simulation	**Domain**	**BOI**	**MEZ**	**STG**
Coast	d01	0.59	-0.03	0.71
URB	d01	0.00	-0.04	0.59
Coast_VL	d01	0.47	-0.19	0.61
ERA5	d01	-2.10	1.54	3.12
Coast	d02	0.81	0.00	0.58
URB	d02	0.83	-0.30	0.52
VL	d02	0.49	-0.06	0.80
Coast	d03	1.13	-0.02	0.50
URB	d03	1.12	-0.30	0.50
Coast_VL	d03	0.89	0.07	0.97

Table 3.5: BIAS of daily two-metre minimum temperature for the different simulations and domains, referred to the selected locations.

benefit is introduced by the additional nestings beyond the first one, except for isolated cases. However, the downscaling itself is effective in reducing summer temperature BIAS across all stations, both in the urban and rural contexts. This was already evident in the tests over a single week seen in Section 3.1, and it is confirmed for this longer time range. The BIAS reduction seems to be especially evident for the daily minimum two-metre temperature BIAS, while the improvement is only marginal for the maximum temperature.

3.3 Tests on a single member of the seasonal ensemble

Given the results in Sections 3.1 and 3.2 and its higher stability, the *Coast* configuration is deemed adequate to downscale the seasonal forecast. As anticipated in Section 2.4, this time the three nested domains have a grid spacing of 27 km, 9 km and 3 km. The seasonal forecast is initialized at the beginning of June, while the downscaling starts on June 2nd at midnight, the first time in the dataset with all variable fields available. The range considered ends on the last day of August. The scores are computed as in Section 3.1, but I only consider the values every 12 hours, at noon and midnight. A direct hourly comparison would not have been possible since the non-downscaled seasonal forecast has the data only available on a sub-daily basis, as explained in Section 2.1. On a positive note, this reduces the computational time. The same data is used to produce the time series, for which a daily statistic is computed by averaging the midday and midnight values. As a reference, the same statistic is computed for the observations coming from the in-situ weather station and the value of the closest point in the gridded observation dataset E-OBS, which characteristics are described in Section 2.1.

I took the control member of the ensemble seasonal forecast, which is arbitrary, as the scope of this test is to assess the quality of the downscaling process itself. The dataset is composed according to what is indicated in Section 2.2 for option A. The scores are computed for the same three stations of Section 3.2. As a comparison, the scores from both ERA5 reanalysis and the non-downscaled non-perturbed member of the seasonal forecast are also considered.

As expected, the seasonal forecast underperforms in terms of scores compared to the ERA5 reanalysis. Interestingly, the dynamical downscaling, forced as described in Section 2.2, provides a clear added value in terms of BIAS, with a reduction that makes it lower than the ERA5 one. While the first two nestings appreciably reduce the temperature BIAS, the difference with the introduction of the third one is limited, as shown in Table 3.6. An advantage in terms of BIAS by the first nesting is also seen in relative humidity BIAS, shown in Table 3.7. This is coherent with the results observed with the ERA5 downscaling in Sections 3.1 and 3.2, which show little to no improvement below the 9 km grid spacing either.

The first nesting is also helpful in marginally reducing MAE, but not below the values for ERA5.

Relative humidity correlation is marginally reduced by the downscaling and ranges from 0.4 to 0.5 across the available stations. Temperature correlation is not affected significantly by the downscaling and remains significantly lower than the ERA5 one.

Simulation	**Domain**	**BOI**	**MEZ**	**STG**
0	d01	-3.67	-1.44	-1.00
ERA5	d01	-1.92	0.46	0.74
SEAS	d01	-4.99	-2.25	-2.91
0	d02	-0.52	-0.96	-0.53
0	d03	-0.50	-0.87	-0.67

Table 3.6: BIAS of two-metre temperature for the different simulations and domains, referred to the selected locations.

Time series

To help frame the scores, I will also present the time series for the mentioned locations. As a reference, also ERA5 and the values from the interpolated observational dataset E-OBS are shown. The values shown in the plots are computed as done in Section 3.2.

Simulation	Domain	BOU	BOI	MEZ
0	d01	5.09	3.86	-4.88
ERA5	d01	14.85	13.46	2.93
SEAS	d01	21.59	20.26	12.85
0	d02	-8.06	-9.29	-5.63
0	d03	-8.01	-9.25	-5.88

Table 3.7: BIAS of two-metre relative humidity for the different simulations and domains, referred to the selected locations.

Looking at Bologna Idrografico, a difference is clear between the months of June, shown in Figure 3.8, and August, Figure 3.9. In the former, there is a clear difference in the shape of the curve compared to its non-downscaled counterpart. This translates into a better forecast in the second half of the month, where the downscaling pushes the forecast closer to the observed values. The original seasonal forecast misses the onset of the heatwave and maintains similar values to the first half of the month.

If we look at the same plot for August, which represents lead time three for the forecast, the situation is very much different. While the downscaling can reduce the negative BIAS because it increases the temperature values on average, there is no clear benefit in terms of the quality of the forecast, as it still does not follow the observed trends. This is coherent with the limited MAE enhancement.

The station data shows temperatures that are consistently higher in value than what both ERA5 and E-OBS indicate. This is likely due to the coarseness of these datasets which are unable to identify the presence of the urban area and its consequences in terms of temperature.

In the rural locations the general trends are very similar to the urban ones. One key difference is the fact that the station observations in all three rural locations fall within the 90th percentile range of the E-OBS dataset. As a representative example for June, the trends from San Pietro Capofiume (SPT) are shown in Figure 3.10. The application of the downscaling introduces a small correction to the original seasonal forecast. This was already evident in terms of the scores, and it suggests the lower impact of a higher resolution and a better soil use representation when modelling temperature outside the urban context. This aspect is further investigated in Chapter 4.

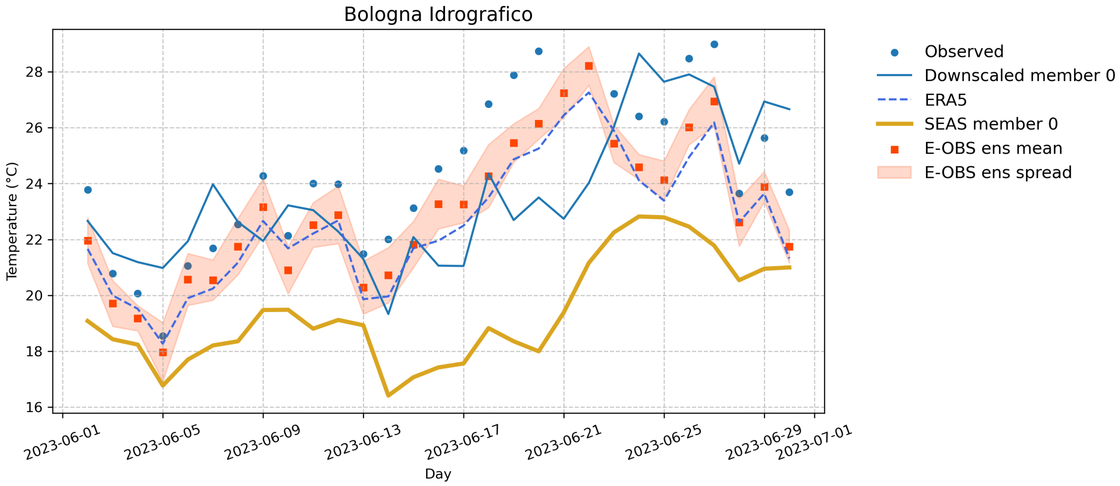

Figure 3.8: Daily values for two-metre temperature for June in Bologna Idrografico. The dots represent the local weather station data. The squares and the shading represent the E-OBS dataset mean and spread respectively and are an alternative observational reference. The thick yellow line is the non-downscaled first member of the seasonal forecast, while the thinner one is the same member after the application of the three steps of downscaling.

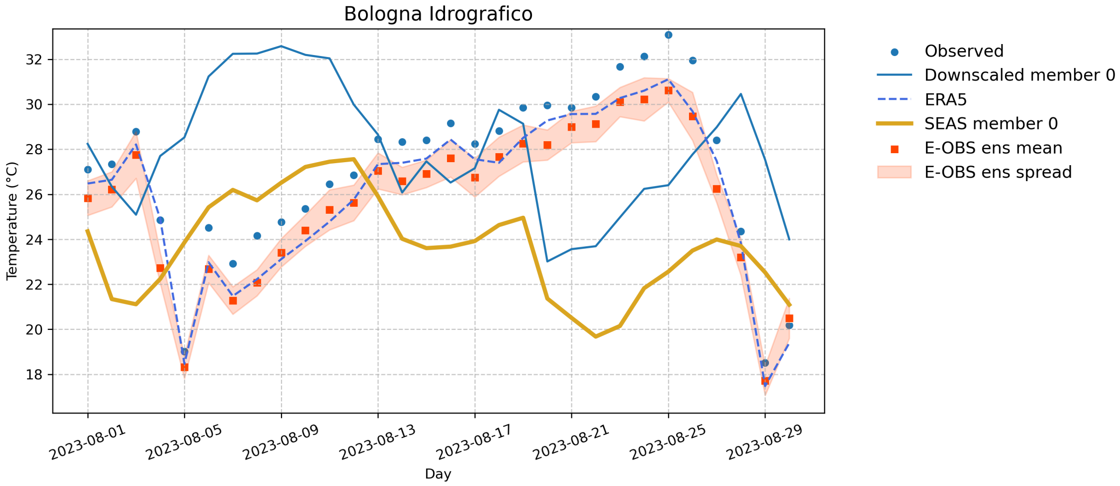

Figure 3.9: Daily values for two-metre temperature for August in Bologna Idrografico. The appearance of the elements in this figure matches that of Figure 3.8.

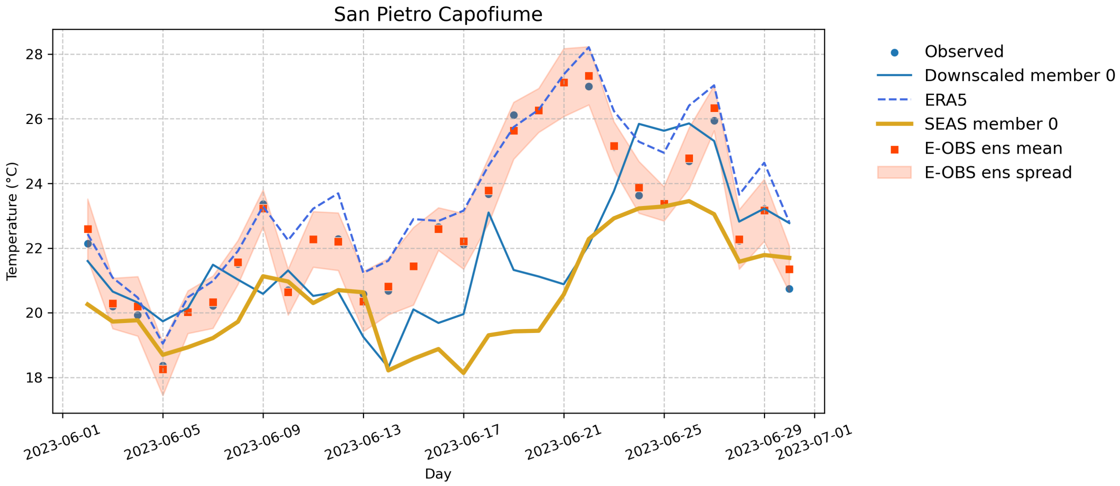

Figure 3.10: Daily values for two-metre temperature for June in San Pietro Capofiume. The appearance of the elements in this figure matches that of Figure 3.8

3.4 Sea surface temperature update

The simulations which scores are shown in Sections 3.1, 3.2 and 3.3 do not update the SST, as is the default in the WRF model. This has the advantage of a lower computational cost. However, it affects the two-metre temperature.

To evaluate the magnitude of the impact on the surface temperature simulation, I took as a reference the ERA5 reanalysis data from 1993 to 2016 and computed the mean summer two-metre temperature field across these years. This allows the computation of anomaly fields for the current year for the different products I want to compare.

Firstly, I present what can be considered the reference anomaly for 2023, namely the ERA5 reanalysis summer temperature anomaly, shown in Figure 3.11. It shows the warm anomaly over most of Europe that characterizes the year.

In Figure 3.12, I instead show the anomaly of the non-downscaled seasonal forecast that I use to initialize the WRF model. Already in this forecast there is a significant difference with ERA5. The highest positive anomalies are observed in the continental areas of the Iberian Peninsula and the mountainous regions of Italy and Morocco. The most negative anomalies are located in the Anatolian peninsula. These areas are dominated by the presence of orography, which oversimplification in the global model may explain the excessive anomalies, but additional tests would be needed to confirm this with certainty. Since the downscaling in heavily influenced by the driving dataset I should expect its anomaly to similarly differ from the ERA5 one.

Figures 3.13 and 3.14 show the mean two-metre temperature anomaly, first without the SST update and then with it active, respectively. Maintaining the SST condition as the initial one leads to an underestimation of the SST in most areas, which in turn produces a lower two-metre temperature than expected over those areas. The impact on land two-metre temperature is less significant. While the difference is clear over the SST and the impact on coastal areas is likely high, the situation is less clear inland. The impact in the set of reference locations in and around the city of Bologna is evaluated in Section 3.4.

While the comparison with the original non-downscaled forecast is imperfect, as the fields have a different resolution, it is still possible to notice how the higher anomalies visible in Figure 3.12 are smoothed out. The update of the SST, which allows taking advantage of the evolving SST boundary conditions provided by the driving dataset, gives an output that is much more similar to the original global seasonal forecast as it reproduces all the main patterns.

To get a first estimate of the impact of this change over the locations of interest, I computed BIAS and MAE over the three summer months. This way I can have reference scores computed in the same way as the other simulations tested in this Chapter. They are shown in Table 3.8 in light blue, next to the corresponding score of the simulation with no SST update. In all instances, there is a reduction of BIAS, with a more significant benefit brought by the finer nestings. I also show the MAE comparison in Table 3.9, for which the relative difference between the two setups is smaller. There is also less difference in the scores between the chosen locations.

Simulation	**Domain**	**BOI**	**MEZ**	**STG**
Member 0	d01	-3.67	-1.44	-1.00
Member 0	d01	-3.25	-0.88	-0.47
ERA5	d01	-1.92	0.46	0.74
SEAS	d01	-4.99	-2.25	-2.91
Member 0	d02	-0.52	-0.96	-0.53
Member 0	d02	-0.03	-0.28	0.02
Member 0	d03	-0.50	-0.87	-0.67
Member 0	d03	-0.03	-0.30	-0.06

Table 3.8: BIAS of two-metre temperature for the different simulations and domains, referred to the selected locations. The rows corresponding to the simulation that updates the SST are highlighted in light blue.

Coherently with what can be inferred from Figures 3.13 and 3.14, the impact of the SST update is marginal yet positive around the city of Bologna, the area on which this thesis focuses. The more realistic SST behaviour increases the temperature and reduces the overall BIAS in all instances. Even the rural stations show a modest BIAS reduction,

Simulation	**Domain**	**BOI**	**MEZ**	**STG**
Member 0	d01	4.99	3.72	3.67
Member 0	d01	4.65	3.53	3.46
ERA5	d01	2.13	1.16	1.65
SEAS	d01	5.76	3.99	4.25
Member 0	d02	3.78	3.79	3.68
Member 0	d02	3.69	3.51	3.56
Member 0	d03	3.76	3.77	3.83
Member 0	d03	3.63	3.52	3.47

Table 3.9: MAE of two-metre temperature for the different simulations and domains, referred to the selected locations. The rows corresponding to the simulation that updates the SST are highlighted in light blue.

with a better score associated to the finer nestings compared to the ERA5 one. This is observed for the urban locations even without the SST update, as the ERA5 reanalysis is likely unable to account for the presence of the city, thus performing significantly worse.

The same comparison is repeated with option B dataset, as defined in Section 2.2. The results are now shown for brevity as they are nearly equivalent and all conclusions drawn for option A are equally valid. Therefore, the SST update is kept active for the final downscaling setup, which results are shown in Section 4.3.

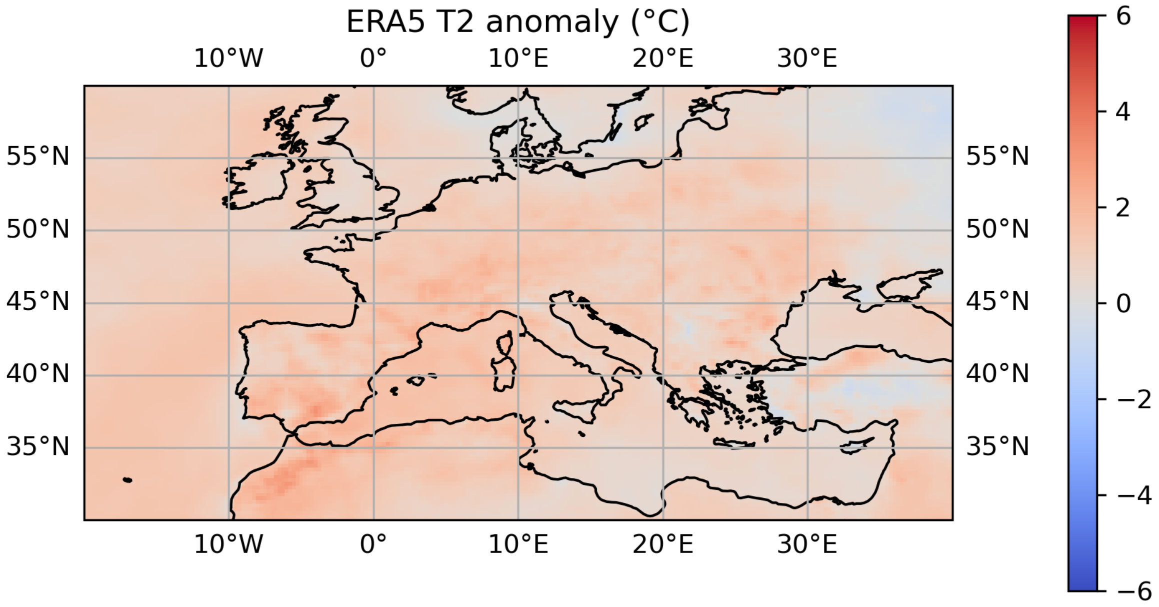

Figure 3.11: Two-metre temperature anomalies of ERA5 2023 reanalysis. The anomaly is positive on most of the domain. This is expected as summer 2023 is chosen for this thesis due to being warmer than usual.

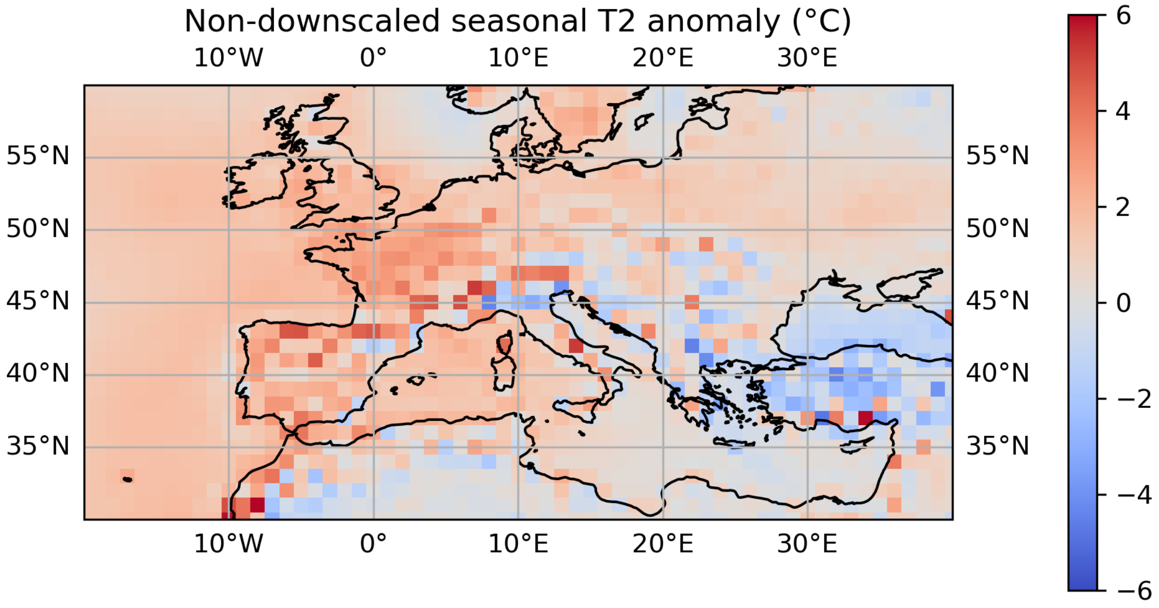

Figure 3.12: Two-metre temperature anomalies of non-downscaled seasonal forecast (member 0) for 2023. The seasonal forecast presents areas with very pronounced anomalies, especially in mountainous areas.

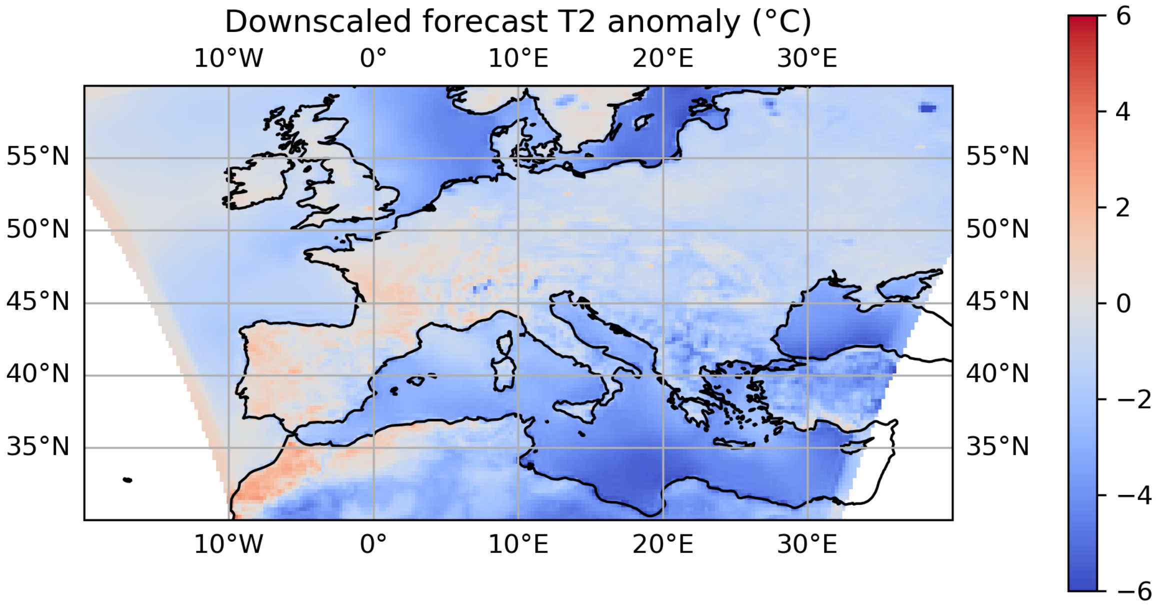

Figure 3.13: Two-metre temperature anomalies of downscaled forecast without SST update (member 0) for 2023. The constant SST leads to an overall underestimation of the two-metre temperature over the Atlantic Ocean and the Mediterranean basin.

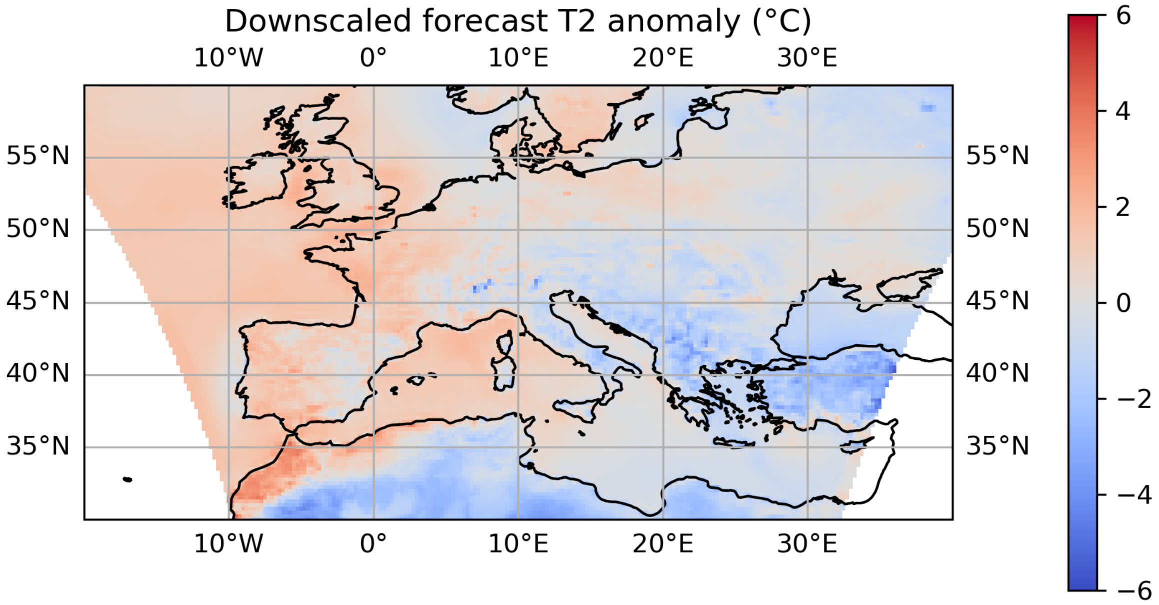

Figure 3.14: Two-metre temperature anomalies of downscaled forecast with SST update (member 0) for 2023. Updating the SST removes the negative anomalies that were present in Figure 3.13 but not in Figure 3.12.

Chapter 4

Results

Given the results presented in Chapter 3, the chosen setup among the ones described in Section 2.4 is now used to downscale multiple members of the ECMWF ensemble forecast. More specifically, the deterministic control run together with the first four perturbed members. The initialization dataset is constructed according to the two alternative options presented in Section 2.2. The references are the observations at each available ground station for the period from June to August 2023.

The results are evaluated on a monthly basis, as explained in Section 2.7.1, they should not be intended as an attempt to estimate the actual model performance for the specified locations, but more as a tool to compare the different adjusted or downscaled forecasts against a common observational reference.

Coarsely-gridded products like E-OBS or ERA5 reanalysis are unable to account for the specific characteristics of local features, like the presence of a city or a particularly narrow valley. This is also true for the set of locations studied in this work, especially the urban ones, as seen in Chapters A and 3. For this reason, in the context of developing a downscaling methodology that can be applied to heatwave forecasting, the comparison with ERA5 scores is also considered.

As an alternative approach, two less costly BIAS correction methods described in Section 2.6 are also introduced. They are both applied to the ensemble mean of the non-downscaled seasonal forecast.

4.1 Preliminary tests - Option A

Initially, I consider the downscaling of the seasonal forecast initialized on June 1st, with the dataset built as explained for Option A in Section 2.2. This means the missing fields are replaced by their constant initial state provided by ERA5 reanalysis.

Let us begin by considering monthly-averaged two-metre temperature BIAS and MAE for a set of locations with available in-situ observations. As it is shown in Figure 4.1, the BIAS for the non-downscaled seasonal forecast is always negative and much smaller than the ERA5 reanalysis one. ERA5 performs better in the rural stations, meaning the BIAS is closer to zero.

The application of the dynamical downscaling generally improves the BIAS, as expected given the test results presented in Section 3.3, referred to the ensemble mean of the downscaled ensemble members for each of the nestings. The average value across all station and months passes from -3.35 of the non-downscaled seasonal forecast ensemble mean to -0.37 after the downscaling process. The positive benefit given by the downscaling is especially clear in the available urban stations, where the scores were initially worse than their rural counterparts. There, after the downscaling, the ensemble means of the two finer nestings outperform even the ERA5 reanalysis. While the improvement when passing from *d01* 27 km nesting to the successive *d02* 9 km is evident, the difference from the latter and *d03* 3 km is barely visible.

The impact of the mean bias statistical correction is instead very marginal and even leads to a slight worsening of the scores in the case of the station of Bologna Asinelli (BOA). The more sophisticated MVA is marginally better than its simpler counterpart in most cases. During June and July, the performance is generally close to the one of the coarser domain ensemble mean. In the rural location during August this correction shifts the BIAS to positive values. In the case of Sant'Agata station (STG), the scores is still close to that of ERA5. The urban locations experience a shift of similar magnitude, which brings the score closer to zero, with a performance that is comparable in these cases with the one of the ensemble mean of the downscaling to the finer nestings *d02* and *d03*.

Similar conclusions can be drawn from a further comparison using the MAE. In fact, in Figure 4.2, it is evident the potential added value of the dynamical downscaling. Using again a simple average across all stations and months, the value for the non-downscaled forecast ensemble mean is 3.35 and is reduced to 0.40 after the downscaling to 3 km resolution.

The urban stations show the biggest reduction in MAE, with the scores of the ensemble mean of the second and third nestings being lower than the ones for the non-downscaled seasonal forecast and the ERA5 reanalysis. In the rural locations in which the score was already lower, the dynamical downscaling benefit is lower. It is still able to bring it to a value comparable to ERA5 reanalysis, and can even outperform it, as is the case in June and August. This also means getting more similar scores across the set of stations after the downscaling, regardless of the nature of their location.

As for the MVA forecast, while performing generally better than the simpler correction, it still does not match in most of the cases the score achieved by the coarser nesting ensemble mean. The biggest leap is observed in the urban locations in the month of August. One exception is the score during the same month in the rural location, where the performance is instead worse than what can be obtained with the simpler correction method.

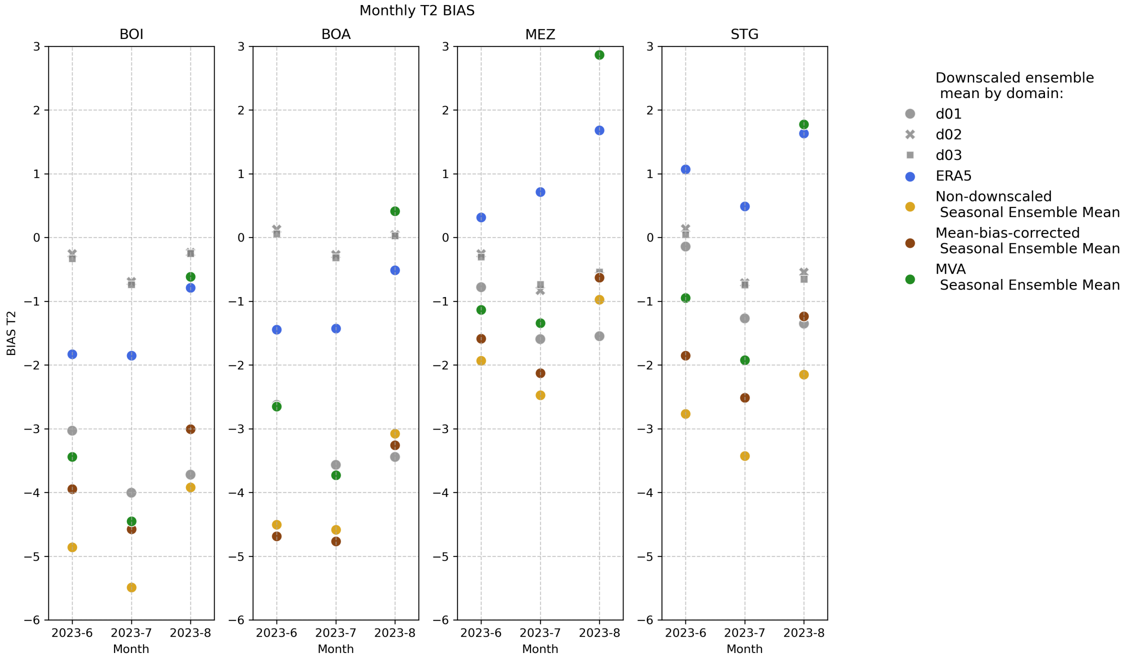

Figure 4.1: Monthly-averaged two-metre temperature BIAS for each of the available locations, using option A. Each grey symbol represents a different domain of the downscaled simulation of which the ensemble mean is taken. The coloured dots represent instead the scores non-downscaled seasonal forecast ensemble mean, its statistical corrections and the one for the ERA5 reanalysis.

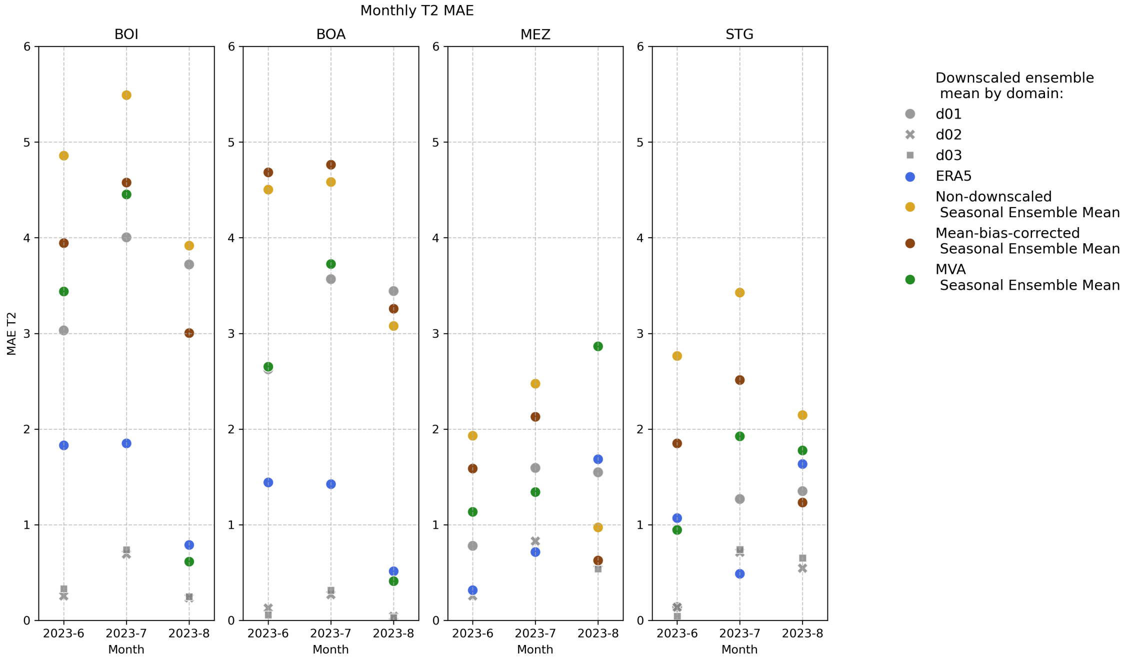

Figure 4.2: Monthly-averaged two-metre temperature MAE for each of the available locations, using option A. The appearance of the elements matches that of Figure 4.1.

To get an idea of the role that an ensemble of downscaled forecasts can play, I also present the time series of the downscaled members. Specifically, the first five members of the ensemble are shown together with the non-downscaled seasonal ensemble mean. The values are relative to the finer nesting *d03*. The daily values shown in the time series are computed as indicated in Section 3.2.

The sample of members to which the downscaling is applied is the result of a random choice, thus rendering the comparison only partially indicative of the potential benefits. However, it allows to have a first estimate of what the behaviour of different members can look like against the observational references and the full-ensemble mean. The bias-corrected forecast are also shown for further comparison.

The month of June is shown in Figure 4.3. While the best-performing member changes depending on the day, the small subset of members mostly follows the general behaviour of the observational reference. This highlights the potential improvement of the forecast quality if the subset of members is chosen properly. The total ensemble mean of the non-downscaled forecast is not surprisingly mostly flat, only showing a mild increase towards the end of the month. However, even considering the standard deviation range the observed temperature values are well above, remarking the importance of member selection. In the other urban stations, the behaviour of the simulations is similar.

In all cases, the effect of the simple mean bias correction, shown in Figure 4.4, is

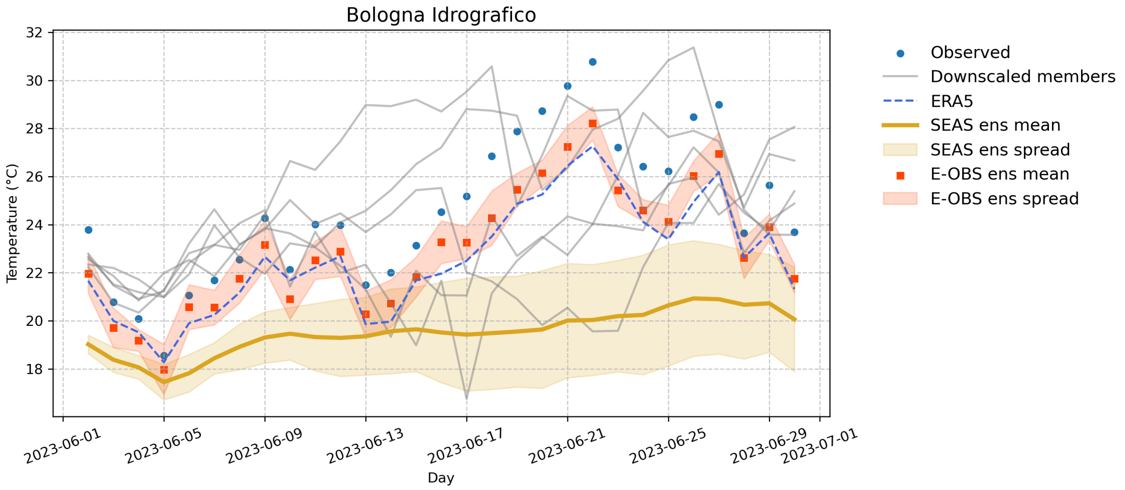

Figure 4.3: Daily values of two-metre temperature for Bologna Idrografico in June, option A. The statistic shown is the one first described in Section 3.2. The appearance of the elements in this figure matches that of Figure 3.8. Here the multiple downscaled members are represented by light grey lines.

a minimal shifting of the trend, which is insufficient to bring the forecast close to the observations in a satisfactory way. The MVA forecast behaves similarly to the single downscaled members. While improving the representation at the beginning and the end of the month, it still does not reproduce the peak intensity of the heatwave, which is crucial when dealing with these extreme events.

Moving to a rural case, in Figure 4.5 the behaviour of multiple downscaled members is shown for the station of Sant'Agata (STG) during June. The ensemble mean is closer to the downscaled members which meander just above it. This time, the members lay within the ensemble standard deviation range for longer. This suggests a lesser difference between the original members of the ensemble and the newly downscaled ones, compared to what is observed for the urban station in Figure 4.3.

4.2 Preliminary tests - Option B

In this paragraph, I present the results related to the alternative preparation of the initialization dataset, option B. As anticipated in Section 2.2, a corrected ERA5 climatology is taken for the missing fields. The reference seasonal forecast is the one initialized in May, and the downscaling is now performed from lead time 1.

In fact, the introduction of the downscaling provides a marginal improvement in the scores in the selected locations, with minor differences between the second (*d01*) and third (*d03*) nestings. The average BIAS across the available stations and months is

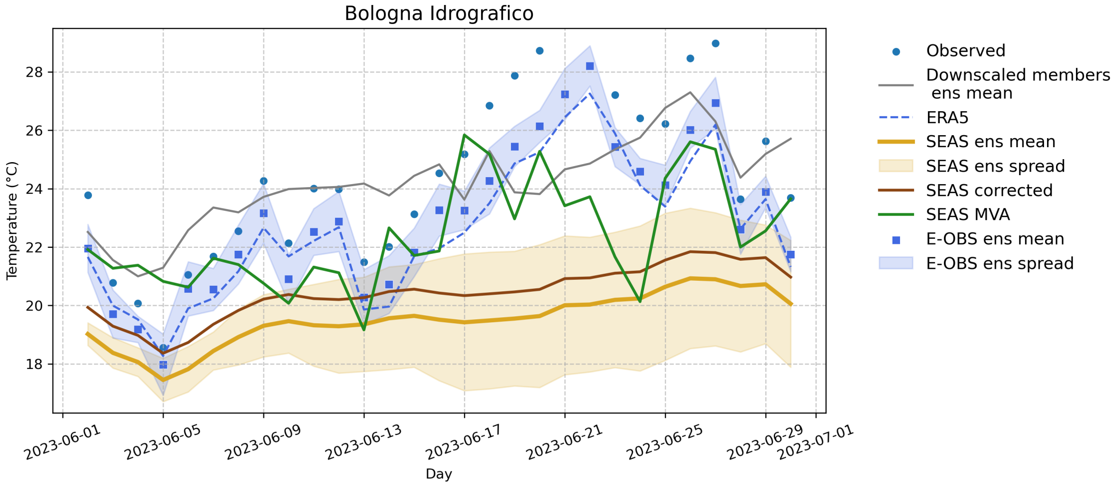

Figure 4.4: Daily values of two-metre temperature for Bologna Idrografico in June, option A. The statistic shown is the one first described in Section 3.2. The appearance of the elements in this figure resembles that of Figure 3.8, with the addition of further continuous lines representing the mean-bias corrected forecast and the MVA one.

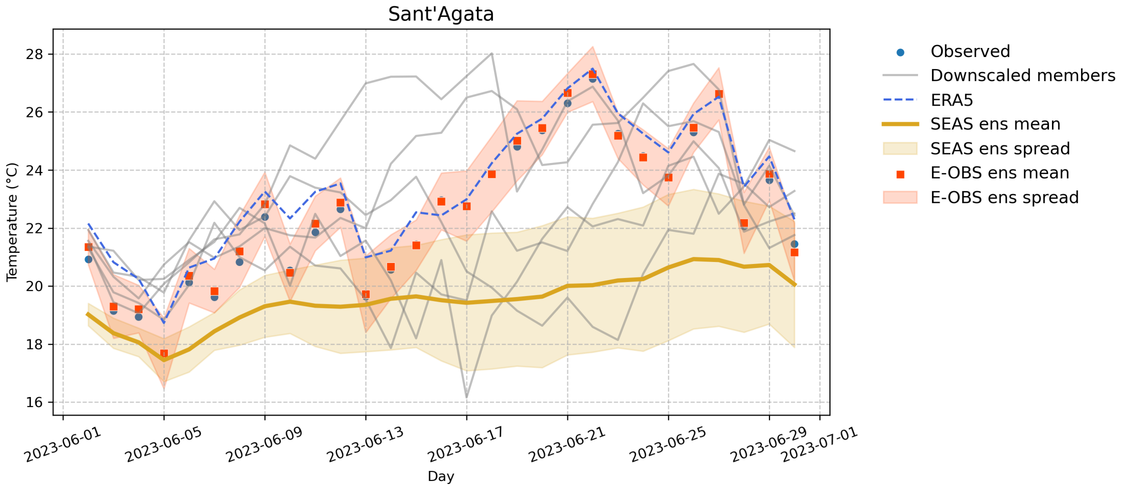

Figure 4.5: Daily values of two-metre temperature for Sant'Agata in June, option A. The appearance of the elements in this figure matches that of Figure 4.3

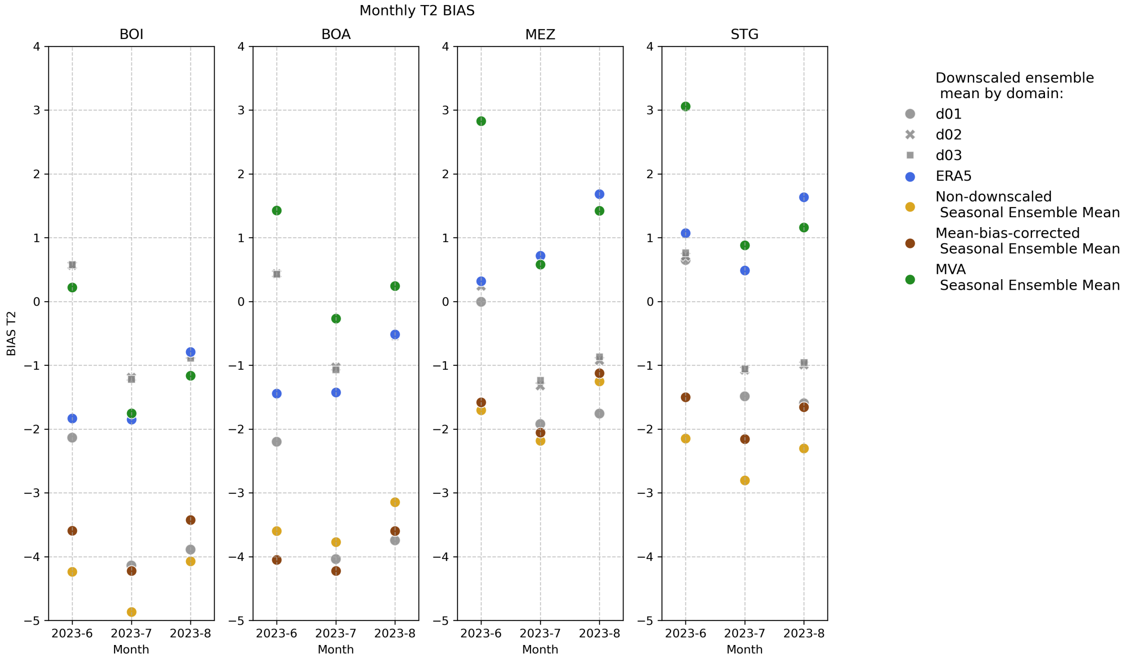

Figure 4.6: Monthly-averaged two-metre temperature BIAS for each of the available locations, using option B. The appearance of the elements matches that of Figure 4.1.

reduced from −3.01 to −0.48 after the downscaling to 3 km resolution. The average MAE is decreased from 3.01 to 0.83. The greatest difference with the non-downscaled seasonal forecast is once again seen in the earlier month and in the urban stations. Therefore, while not being directly comparable with the seasonal downscaling based on Option A, as it is driven by a different forecast, the effects on the scores are indeed similar. During the first month, the BIAS moves from being negative to positive in sign, as shown in Figure 4.6. At the same time, there is a decrease in MAE, shown in Figure 4.7.

In this instance, the simple mean bias correction has a very small impact on the forecast, as seen in Section 4.1. MVA tends to perform well in terms of BIAS in the urban stations, resulting in better values than the simpler correction that renders it comparable with the scores of the finer nestings ensemble mean and those of ERA5. In rural locations, it is instead the simpler correction that leads to a better result among the two, as the pronounced shift leads to a positive BIAS. It can be even higher in magnitude than the initial negative BIAS of the non-downscaled seasonal forecast, as it happens in the case of June in rural locations. The erratic effect it has in this instance renders it inadequate.

Moving to the MAE, most of the conclusions remain valid, as a better BIAS is reflected in a smaller MAE. The scores are shown in Figure 4.7.

The MVA performs well in the urban context in all months. In the rural location,

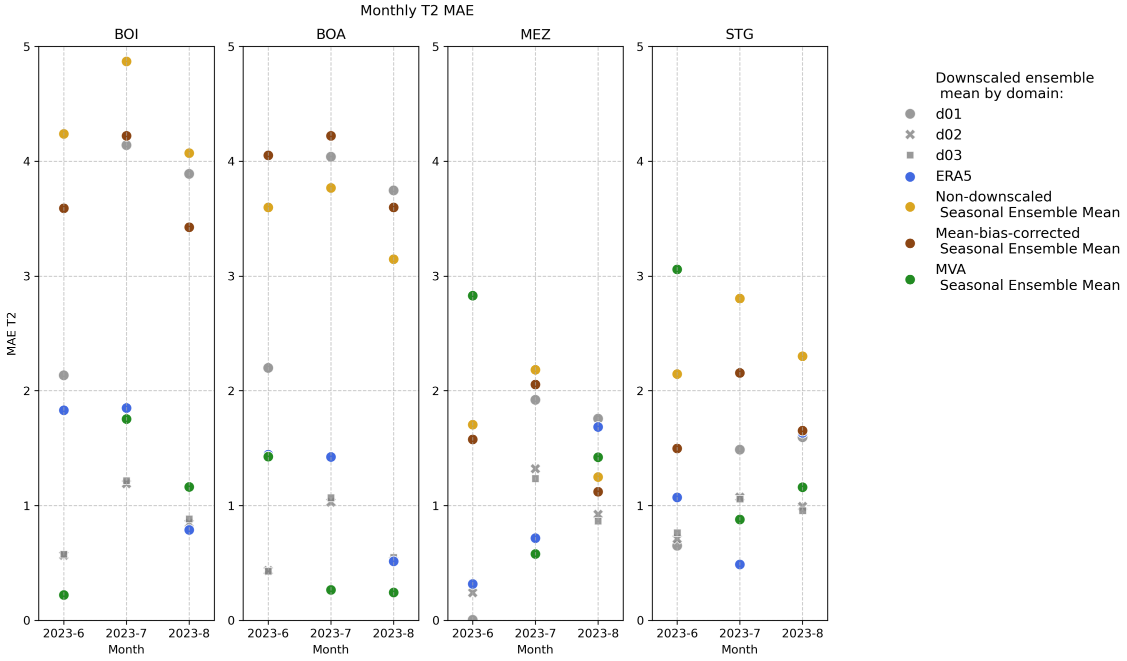

Figure 4.7: Monthly-averaged two-metre temperature MAE for each of the available locations, using option B. The appearance of the elements matches that of Figure 4.1.

its poor performance for the month of June is confirmed, with a higher MAE than the non-downscaled seasonal forecast. For the other months, the scores are instead improved by the correction. This is coherent with what is observed in Figure 4.6, as the BIAS is smaller in magnitude than the non-downscaled seasonal forecast, even if it has its sign changed.

Let us now consider the time series of two-metre temperature. Figure 4.8 shows the month of June for the station of Bologna Idrografico (BOI). In agreement with the scores, the downscaled ensemble members mostly overestimate the observed temperature, especially at the beginning of the month. The MVA forecast, shown in Figure 4.11 has a similar behaviour, as expected given the similar score previously presented.

Moving to the rural station of Sant'Agata (STG), shown in Figure 4.10, which situation is analogous to the one in Mezzolara (MEZ), it is possible to associate the strong positive BIAS of the MVA forecast, in Figure 4.11 with an even more pronounced overestimation of temperature, especially in the first part of the month.

The trend of August is not followed accurately by any of the members. As for June, in the urban context, they are much closer to the observation than they are to the non-downscaled ensemble mean. The MVA forecast remains closer to it, as shown in Figure 4.12.

The worse performance for the rural stations is confirmed by looking at the time series.

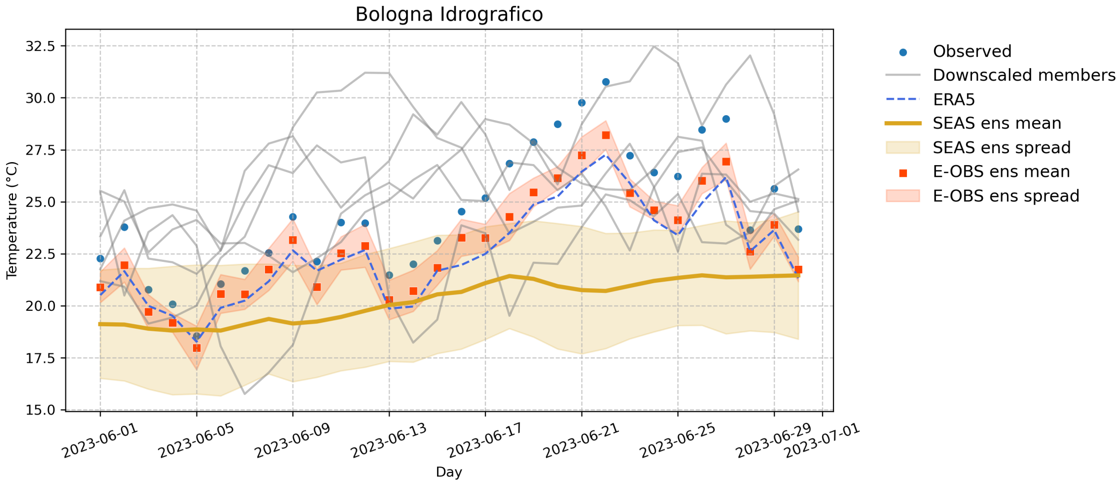

Figure 4.8: Daily values of two-metre temperature for Bologna Idrografico in June, option B. The appearance of the elements in this figure matches that of Figure 4.3.

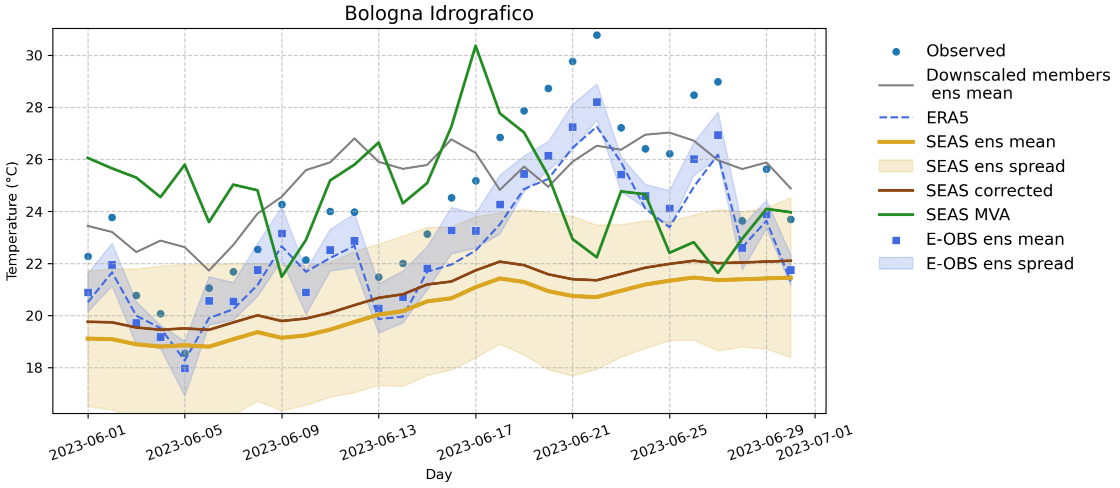

Figure 4.9: Daily values of two-metre temperature for Bologna Idrografico in June, option B. The appearance of the elements in this figure matches that of Figure 4.4.

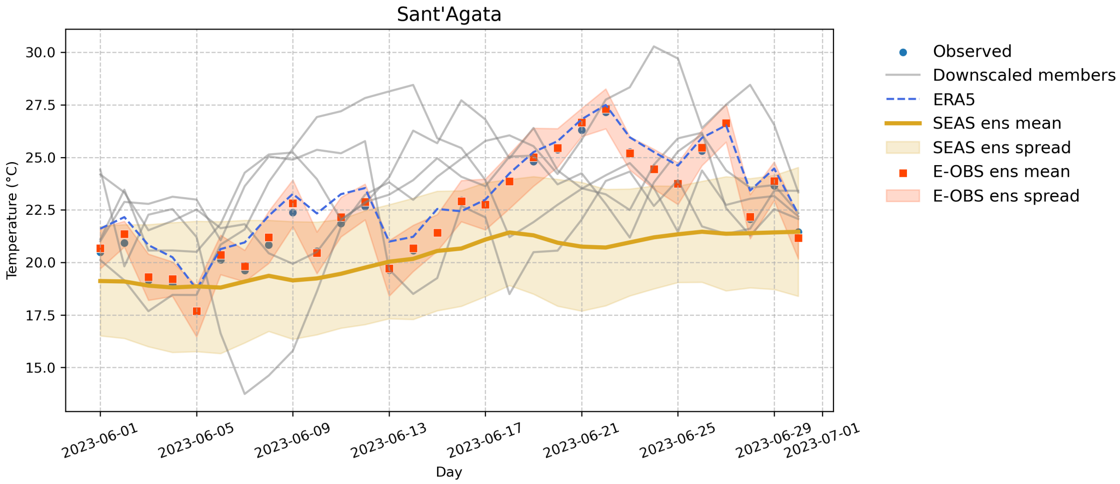

Figure 4.10: Daily values of two-metre temperature for Sant'Agata in June, option B. The appearance of the elements in this figure matches that of Figure 4.3.

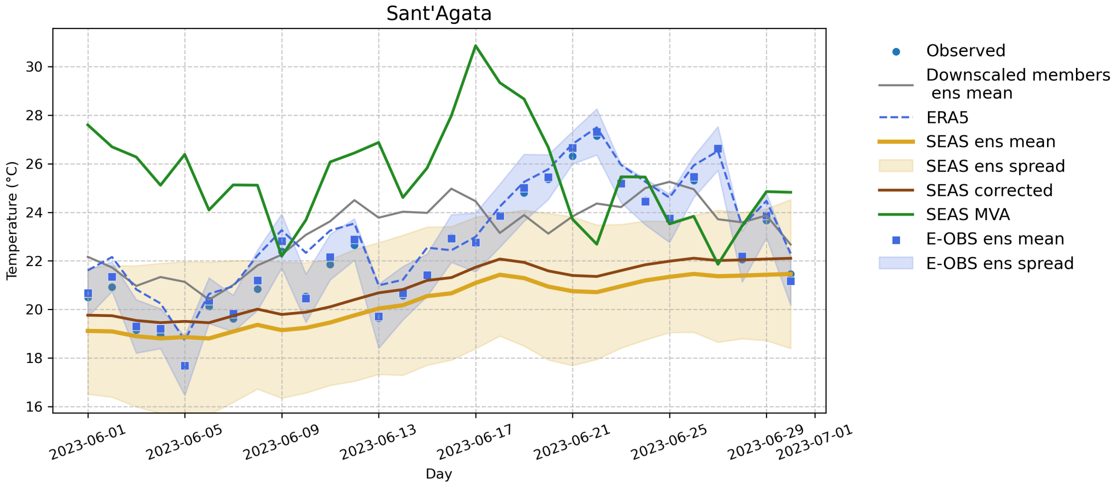

Figure 4.11: Daily values of two-metre temperature for Sant'Agata in June, option B. The appearance of the elements in this figure matches that of Figure 4.4.

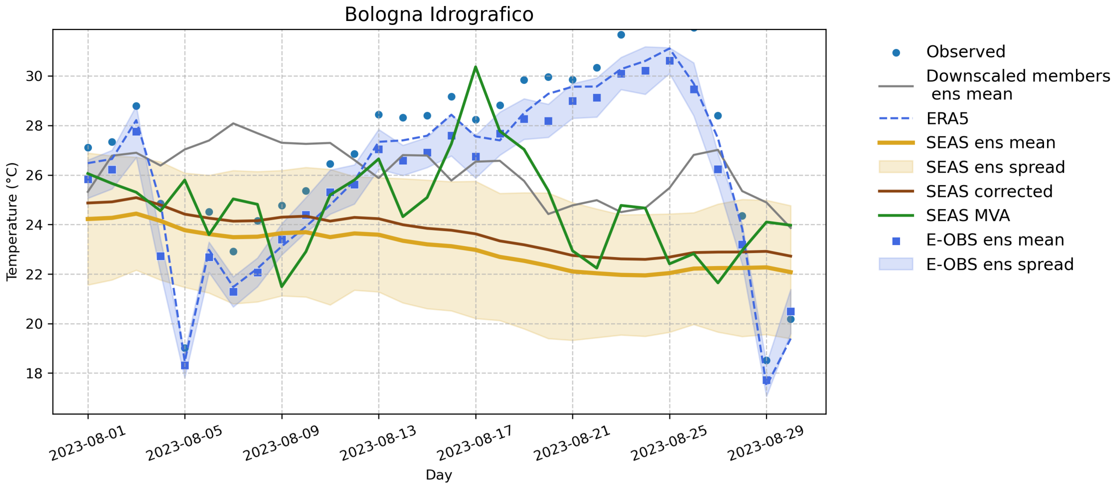

Figure 4.12: Daily values of two-metre temperature for Bologna Idrografico in August, option B. The appearance of the elements in this figure matches that of Figure 4.4.

Taking as a reference the series in Sant'Agata (STG), most members underestimate the surface temperature in the second half of the month, losing the onset of the heatwave, with values closer to the non-downscaled ensemble mean.

In essence, the dynamical downscaling is effective in reducing monthly-averaged two-metre temperature BIAS and MAE within the urban context, while the benefit is smaller in the rural locations, which performance of the original forecast is already better. Therefore, the scores across the different locations are similar after the downscaling.

The performance of the dynamical downscaling based on Option B is similar to what is observed in 4.1 for Option A. However, since the considered simulation is the product of downscaling a seasonal forecast initialized one month prior, one can argue this option works better, although more comprehensive testing would be required to assert it.

The simple mean variance correction only has a minor impact on the forecast. It cannot compete with the other methods as it is incapable of accounting for the specificity of a certain location. The performance of the MVA forecast is generally positive. However, it is disappointing in some cases, cautioning against an unsupervised use of this technique.

4.3 Analysis of the final setup

In this Section, I will assess the monthly performance of the simulation that uses option B as the initialization dataset and introduces the SST update. This final configuration, which characteristics are detailed in Section 2.4, is chosen according to the verification results in Chapter 3. The overview Table 4.1 summarizes the key aspects of the setup.

Category	**Description**
Indicators computed	Two-metre temperature and relative humidity
Temporal resolution of indicators	12 hours
Initial conditions	As per option B
Boundary conditions	As per option B
Domains	3 nested domains
Ensemble members	5 (subsampling)
Parent model resolution	1 degree
Child model resolutions	27 - 9 - 3 km

Table 4.1: Key aspects of the final seasonal forecast downscaling setup.

The impact of updating the SST seems to be low on inland locations, as addressed in Section 3.4.

Similarly to what is done in Sections 4.1 and 4.2, I compute the BIAS and MAE of monthly-averaged two-metre temperature, shown in Figures 4.13 and 4.14 respectively. Considering the scores across the set of stations and the three months, the average BIAS passes from -3.01 of the non-downscaled forecast ensemble mean to 0.15 after the downscaling to 3 km resolution. This low score is also the consequence of compensating positive and negative biases. The average MAE goes from 3.01 to 0.62.

The BIAS in the different cases is not far from the results of the preliminary tests without the continuous SST update, summarized at the end of Section 4.2. The main difference consists of generally higher values, which testifies a tendency to shift towards warmer temperatures. This is expected given the outcomes presented in Section 3.4. It leads to a marginal improvement in July and in the scores of the coarser grid in the urban stations. Both cases were previously affected by a general underestimation of temperature, so they benefit from the increase. The impact is less clear in other instances, with the ones presenting positive BIAS that have it further increased with the SST update.

Considering the MAE, similarly to what is seen in Section 4.2, a distinct improvement is provided by the dynamical downscaling in the urban context. The scores for the downscaled ensemble mean of the two finer nestings *d02* and *d03* are lower or nearly identical to those of ERA5. The difference is smaller for the rural locations, as the non-downscaled forecast already has a lower score, but the downscaling is still beneficial, and the score is similar to the one of the ERA5 reanalysis.

These results are reflected in the time series. More specifically, to further investigate

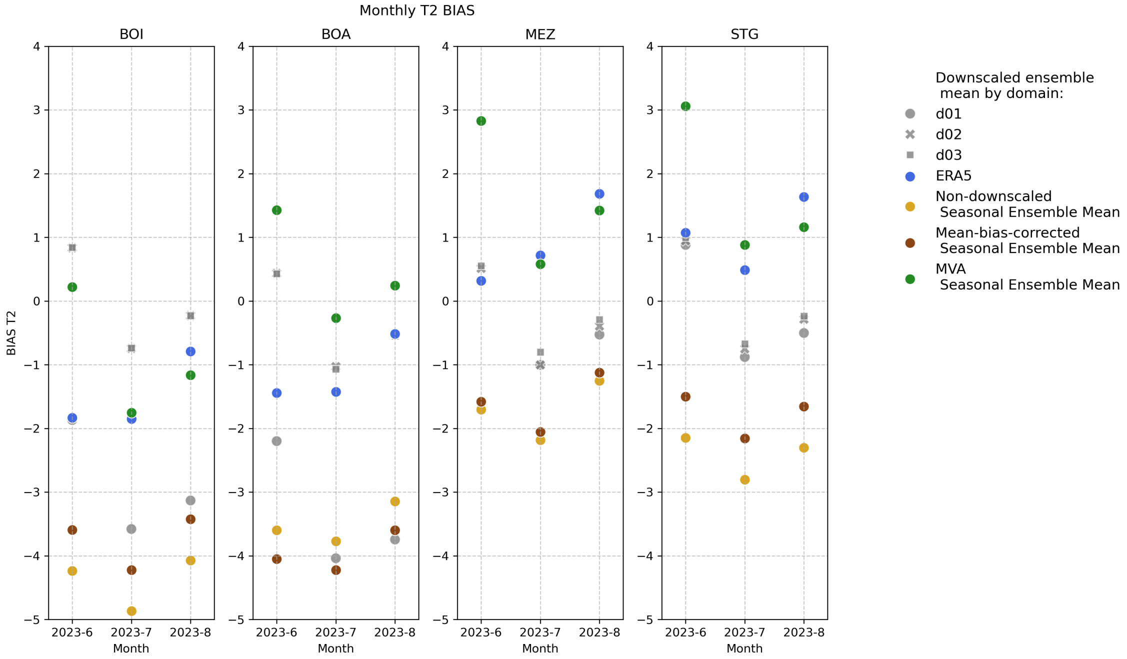

Figure 4.13: Monthly-averaged two-metre temperature BIAS for each location (final setup). The appearance of the elements matches that of Figure 4.1.

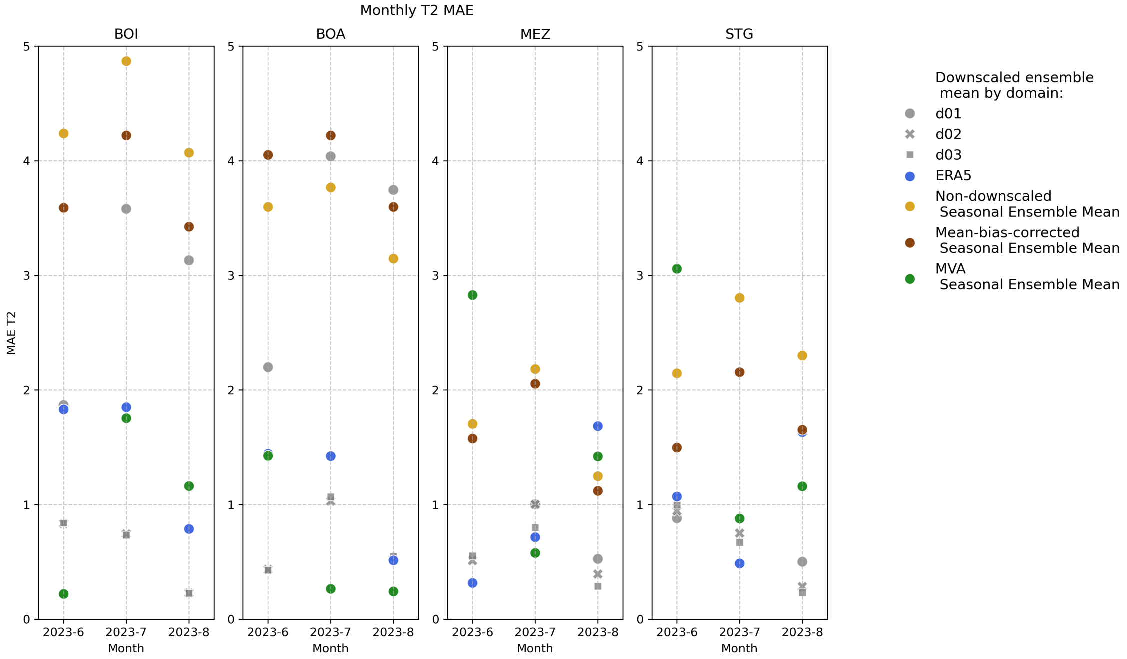

Figure 4.14: Monthly-averaged two-metre temperature MAE for each location (final setup). The appearance of the elements matches that of Figure 4.1.

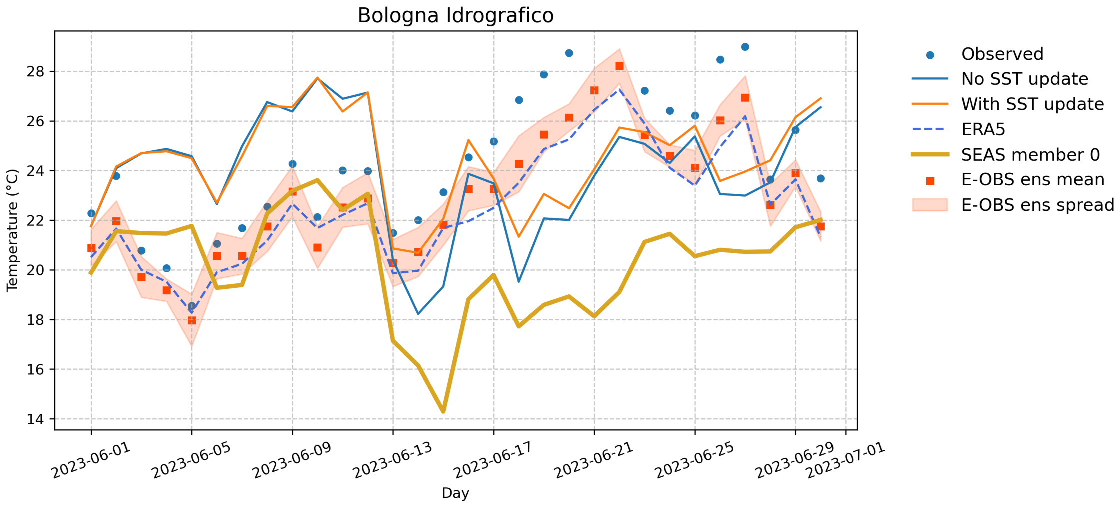

Figure 4.15: Daily values of two-metre temperature for Bologna Idrografico in June. Option B without the SST update is represented by a blue line. The orange line represents the same setup but with the SST update. The appearance of the other elements in this figure matches that of Figure 3.8.

the differences in the daily values introduced with this variation, it is possible to consider the time series for the single members with and without the constant SST.

Since the situation is similar across the locations and months, I chose Bologna Idrografico for June as a representative example. The trend in Figure 4.15 shows how the SST update brings the first member of the simulation marginally towards warmer temperatures, thus explaining the higher BIAS seen in Figure 4.13 and anticipated in Section 3.4. The figure shows how this difference is minimal at the beginning of the month while getting more pronounced in the following week, as the SST temperature increases and distances itself from the initial condition. The magnitude of this difference does not increase further in the following months.

As the differences are quite limited from a visual standpoint and the plots would look nearly identical to the ones shown in Section 4.2, they are omitted for brevity. Refer to this Section also for a discussion of the statistical corrections, as the same initial seasonal forecast is corrected with the same methods.

Chapter 5

Conclusions

In this thesis, I propose, implement and discuss a methodology to downscale seasonal forecasts that uses open-source data with global coverage. More specifically, it uses a combination of seasonal forecast and ERA5 data as initial and boundary conditions for the WRF model. The downscaling is performed up to a resolution of 3 km. The focus is assessing any improvement in two-metre temperature representation compared to the non-downscaled seasonal forecast in specific locations, using in-situ observations as reference.

In the first phase of verification, described in Chapter 3, where ERA5 is downscaled with the different setups, there is a clear BIAS reduction. This is especially true for the night minima, meaning that dynamical downscaling has the potential to address the issue of wrong nighttime temperature representation. This is a key aspect of the well-being of the population during heatwaves, as there is evidence this is the time of the day with the highest UHI amplification Possega et al. 2022. A reduction in monthly-averaged two-metre BIAS and MAE is observed in the selected locations even when downscaling the seasonal forecast. Using the final setup, the average MAE value across all stations and months is lowered from 3.01 to 0.62.

Introducing the sea surface temperature update in WRF is crucial to avoid a cold BIAS over the ocean and the Mediterranean Sea, which can significantly affect coastal areas. In the tested locations, which are further inland, the impact is more marginal. Dynamical downscaling alone cannot be the only way of addressing the presence of BIAS. As highlighted in the season-averaged fields, the two-metre temperature BIAS is only partially corrected with the downscaling.

The statistical correction methods are a way of avoiding the high computational cost of dynamical downscaling. The simpler one is not very effective, as by definition it cannot change the shape of the trend it is correcting. This does not improve the representation of the variability and can only slightly reduce the BIAS. The more complex MVA has the potential to be competitive with dynamical downscaling. However, it can perform very poorly in isolated cases, as shown in Chapter 4, and those must be excluded. This requires further and wider testing as an acceptance threshold needs to be defined.

The testing setup is limited and so is the fraction of data considered in the analysis. For this reason, this thesis can only support speculative conclusions. Nonetheless, I find that the results from this downscaling methodology are encouraging. This is especially true in terms of monthly-averaged two-metre temperature BIAS and MAE reduction in urban locations. Based on the presented results, I recommend completing the simulations to cover the subsampled dataset for the entire dataset hindcast period from 1993 to 2016. These new simulations should be reviewed in light of the residual bias identified in this thesis.

As the simulation proceeds, the inspection of the time series reveals how the downscaling of the forecast is unable to meaningfully change the trend, which largely follows that of the original forecast for the subsequent months of the simulations. This underlines the importance of ensemble subsampling in picking the best-behaving members from the beginning. The subsampling should be process-informed, meaning the selection is guided by documented local predictors, with a well-explained physical connection.

In those cases in which it is deemed applicable, the statistically corrected forecast can be one component of a parallel hybrid forecasting approach, in which the output from statistical and dynamical models are used together. A review of this technique and recent developments is available in Slater et al. 2023.

Given the heatwave impact on society, anticipated in Chapter 1, an impact-based prediction can provide added value to this forecast. This can be done by linking impact assessment models, as suggested by AghaKouchak et al. 2022, which in this case would estimate the impact of forecast extreme temperatures. For instance, the increase in water consumption or the burden on the healthcare system.

Appendix A

The concept of score

The concept of score relies on the definition of an observation (or target) and an assigned probability distribution of the possible outcomes coming from a forecast. With a reliable forecast the observation follows the forecast distribution. In the simpler case of a discretized distribution comprising n classes of events, it is possible to define a scoring rule $S(a, k)$ as a function taking one set a of n assigned probabilities and one possible observed outcome k. The scoring function of two different forecast probability sets, Equation A.1, is defined as the sum of the scoring rule functions of the first set over the n possible types of events, each one multiplied by the corresponding probability of the second set. If they are equal, such quantity is defined as the entropy of that forecast probability set. From the scoring function, it is possible to define the divergence, given by Equation A.2.

$$s(a,b) = \sum_{k \in classes}^{n} S(a,k) b_k \tag{A.1}$$

$$d(a,b) = s(a,b) - s(b,b) \tag{A.2}$$

Given a certain scoring rule, the score is said to be strictly proper if it is zero only when the two sets are equal, and always positive in the other cases. If this holds, it is possible to decompose the expectation value of the score in three parts, as shown in Equation A.3. π_k^a is the conditional probability of getting k as an observation given a being the forecasting scheme and $\overline{\pi}$ is its expectation value and represents the climatology. The entropy of climatology represents how informative climatology is when used as a forecast. The resolution term quantifies the information carried by the forecast. The reliability term indicates the discrepancy between the forecast and the climatology. Further details and a complete demonstration of this decomposition are available in Bröcker 2009.

$$\mathbb{E}[s(a,k)] = s(\overline{\pi}, \overline{\pi}) - \mathbb{E}[d(\overline{\pi}, \pi^a)] + \mathbb{E}[d(a, \pi^a)] \tag{A.3}$$

Appendix B

Statistical calibration methods

In the following chapter, I am presenting with more detail the main existing statistical correction methods, which categories have been anticipated in Section 2.6.

Each of these classes of methods operates based on a different logic, with the more complex techniques focus on addressing one or more specific issues. Depending on the intent, these methods can be applied before the data is fed into a model or after its execution, to post-process the output. Some of these methods are suitable to be used in combination with a decomposition technique to isolate a precise signal within the variability of the object of interest. For instance, Empirical Orthogonal Functions (EOF) have been used to filter out unnecessary noise in Zorita and Von Storch 1999.

The notation will be maintained throughout the Appendix. The overline indicates the temporal mean, the angled parentheses indicate the ensemble mean, and the apostrophe marks the corrected model output. σ indicates the standard deviation of the observations (σ_o) or the model, considering all members of the ensemble and all time (σ_m) or all times of the ensemble mean ($\langle\sigma_m\rangle$). ρ represents the long-term correlation between the ensemble mean and the reference observations.

B.1 Statistical downscaling using MOS and PP

Model Output Statistics (MOS) and Perfect Prognosis (PP) are two possible statistical downscaling approaches. They share a lower computational cost than dynamical downscaling. On the other hand, they both require a time-consuming screening process to identify the most appropriate variables to be used as predictors. The predictor is a large-scale variable showing a physically-backed statistical link with the local predictand of interest. It can be accomplished using either historical data or observations, for which we then need a high enough spatial resolution, as explained in G. T. Diro, Tompkins, and Bi 2012. A reliable characterization of the local climate is necessary to implement them correctly, otherwise they might deteriorate the result (Manzanas, J. M. Gutiérrez, Bhend, Hemri, F. J. Doblas-Reyes, Penabad, et al. 2020). Moreover, their performance depends on the region and model in use.

Model Output Statistics

MOS are trained with predictors taken from the model that is being post-processed. They create a correspondence between past numerical forecasts and past observations which is then used with current numerical forecasts to infer a value for the variables of interest. This means they can only work on a monthly or seasonal basis. They work best when a large dataset of model and observation data is available. By construction, they significantly reduce the bias in the forecast since it automatically removes any systematic bias, as shown in Marzban, Sandgathe, and Kalnay 2006. For this reason, many studies in the past made use of MOS.

When evaluating this technique, it is important to use a cross-validation framework. This avoids the overestimation of skill improvement, as explained in Manzanas, J. Gutiérrez, et al. 2018.

Perfect Prognosis

PP encompasses a variety of techniques which are based on the use of observation data instead of direct model output. Both the predictand and the predictor are observations. A regression model is trained using large-scale predictors, coming from reanalysis data, which are assumed to be a perfect forecast (Maraun 2016). This gives the name to this class of methods.

PP is particularly recommended for those circumstances where the large-scale predictor is significantly better represented than the local predictand, otherwise the improvement following the introduction of this technique is minimal. This should be considered as the regression model training can be expensive, and they may render its benefit disproportionate to the actual resource cost (Manzanas, J. M. Gutiérrez, Bhend, Hemri, F. J. Doblas-Reyes, Torralba, et al. 2019).

The complexity of the possible regression models used ranges from simple linear ones to generalized regressions. Other options are analogue techniques, as in Zorita and Von Storch 1999, which require the existence of a wide catalogue of past conditions to refer to.

B.1.1 Dealing with model imperfections

One further aspect to consider is that during the integration of a model, there is a loss of skill due to both the imperfections in the model and the chaotic nature of the atmosphere associated with the non-linearity of the processes. When a PP or MOS technique is applied the relation between the predictor and the predictand could be built contemporaneously for the two at a certain time. This is the best choice if we assume that the model is perfect. But no model is perfect, and if the rate of loss of information by the model is superior to the intrinsic one from chaos an optimal time lag can be introduced to improve the training, as explained in Marzban, Sandgathe, and Kalnay 2006. The higher the uncertainty of the model is, the longer the time lag has to be. The limiting case is a completely flawed model for which the predictor is taken from its initial condition. Marzban, Sandgathe, and Kalnay 2006 proposed a reanalysis-based approach to estimate the model uncertainty by using a double regression model.

B.2 Bias adjustment techniques

Bias Adjustment (BA) methods are regularly applied to general circulation models to reduce their bias before using them to drive a regional circulation model (Hoffmann et al. 2016). This may be necessary if a bias is discovered for a given region in the host model as it can propagate into the regional model which is dominated by the boundary and initial conditions imposed (Pielke and Wilby 2012). However, when considering whether to perform a scaling, it is important to understand the nature of the observed bias from the climatology. The non-scaled variables may still lead to a more realistic outcome and thus can be preferable for operational seasonal prediction, as recalled in Koster, Mahanama, Yamada, Balsamo, Berg, Boisserie, Dirmeyer, F. J. Doblas-Reyes, et al. 2011.

Bias adjustment methods are not downscaling methods and thus work better in those cases where the output is similar to the observation. In the following sections some alternative approaches are presented. Some of them are simple and are characterized by a lower computational cost, while others are designed to work in conjunction with a dynamical downscaling.

Mean and variance bias correction

Typically, the correction tightly follows the assessment of the model anomaly. It is commonly computed using a lead-time-dependent climatology as a reference, obtained from an available set of hindcasts, as explained in Meehl et al. 2021.

As anticipated in Section 2.6, the mean bias correction techniques are usually based on a linear scaling, like the one in Lenderink, A. Buishand, and Van Deursen 2007, shown in Equation B.1. In this case, the ensemble members are concatenated instead of being averaged directly, and the time averaging operation is followed by a smoothing using a Gaussian filter. An additive term is typically used in the case of temperature correction, and it is based on the difference between long-term monthly mean observed and modelled values (Teutschbein and Seibert 2012).

$$x'_m = x_m + (\overline{x_o} - \overline{x_m}) \tag{B.1}$$

Crochemore, Ramos, and Florian Pappenberger 2016 applied a linear scaling based on a multiplicative factor to the correction of ensemble seasonal precipitation forecasts, which helped improve forecast accuracy. Ghimire, Srinivasan, and Agarwal 2019 suggests that a better precipitation correction can be accomplished with correction factors computed separately on a monthly basis. The factor is defined as the ratio between the observed values and the model ensemble mean, and it is used to rescale the monthly mean values of the precipitation forecast.

Another way of correcting the mean bias of the GCM is the mean shift correction used by Holland et al. 2010. Since then, it has been applied both in climatological studies and in seasonal forecasting. It is referred to with the name of Mean Adjustment (MA), and it is given by Equation B.2.

$$x'_m = (x_m - \overline{\langle x_m \rangle}) + \overline{\langle x_o \rangle} \tag{B.2}$$

However, only correcting the mean bias leaves untouched all other biases. The use of a non-linear scaling allows acting on the variance bias as well, even with a simple power transformation as shown for precipitation in Leander and T. A. Buishand 2007. It has been applied to precipitation correction also by Teutschbein and Seibert 2012. A more direct approach is the Mean and Variance Adjustment (MVA), which equation is presented in Section 2.6.

Xu and Zong-Liang Yang 2012 implementation, where biases are assumed stationary in time, has been found to significantly improve the forecast, even regarding extreme events Xu, Y. Han, and Z. Yang 2019. On the other hand, it may alter the trend of certain variables. Hoffmann et al. 2016 proposed a procedure to circumvent this problem, by removing the trend and adding it back after the correction has been applied.

Quantile-quantile correction

The quantile-quantile correction (Colette, Vautard, and Vrac 2012) is another BA technique based on the correction of the cumulative distribution function towards the reference global one. Also known as Quantile Mapping (QM), these methods alter the shape of the distribution, allowing them to correct biases even in the extremes.

There are different kinds of QM techniques. The different implementations depend on the distribution chosen for the cumulative function. Golian and C. Murphy 2022 evaluated multiple alternatives, while also checking whether it is more beneficial to apply the correction to the individual members of an ensemble of GCM and then take their average or directly to the ensemble mean. It is shown that whichever QM method is applied, the former results in a lower Mean Absolute Error and better correlation. On the other hand, this approach should be avoided when considering the extremes of the distributions, since it tends to neutralize them. In this case, the latter choice can be preferable.

QM needs to be applied carefully as it may introduce spurious precipitation variability and additional bias in spatial gradients, as explained in R. H. White and Toumi 2013. Another aspect worth mentioning is that such correction does not maintain the inter-variable dependencies since it acts on the distributions without considering the connection between the different variables. An improvement on this matter may come from the introduction of consistency constraints, on which you can read more in Section B.2.

One common QM choice is the Empirical Quantile Mapping (EQM) which adjusts the percentiles of an empirical cumulative function. However, it can perform worse when compared to MVA, at least for some observational references (Manzanas, J. M. Gutiérrez, Bhend, Hemri, F. J. Doblas-Reyes, Torralba, et al. 2019). It is also significantly more resource-demanding. Nonetheless, it can still be a valuable tool for extreme adjustments and for all those indicators based on thresholds. In another study by Crochemore, Ramos, and Florian Pappenberger 2016, this technique was effective in improving forecast reliability.

Depending on the variable considered, parametric distributions can be used as well. For instance, a Gaussian distribution or a Gamma distribution can be used for precipitation and temperature data respectively (Manzanas, J. M. Gutiérrez, Bhend, Hemri, F. J. Doblas-Reyes, Penabad, et al. 2020).

Delta Mapping (DM) is another possible QM method, which has been shown to outperform the QM methods based on parametric distributions in some regional studies (Mendez et al. 2020). It operates while maintaining the relative changes in quantiles.

Bias correction with consistency constraint

To avoid breaking the dependencies between different variables it is possible to introduce bias correction methods with consistency constraints, which may impose for instance the hydrostatic equilibrium or the geostrophic balance. Meyer and Jin 2016 proposed a multistep procedure, where only a few variables are corrected, and then the others are computed from them through known relationships.

A further scheme proposed by Hernández-Díaz et al. 2017 introduces a procedure that can maintain the inter-variable dependencies while also performing a three-step dynamical downscaling. It uses an atmosphere-ocean coupled GCM in which sea surface temperature and sea-ice mean biases are corrected, providing a higher added value in those regions highly affected by these variables. Then a second model is introduced, an atmosphere-only GCM, forced by the previous one and formulated in a way that retains the wanted dependencies. The last RCM model is driven by the second. Significant degradation due to the bias is still possible in the upper variables far from the direct influence of the sea. As a consequence, it may not work properly in those cases in which the sea surface temperature is strongly influenced by atmospheric conditions, as explained in Xu, Y. Han, and Z. Yang 2019.

In many cases, the additional correction given by the constraints can be negligible. Given the high computational cost of this approach, it may not be the best choice even considering its higher physical coherence. The correction might be so small compared to intrinsic variability that it can be smoothed out anyway by the RCM integration, as shown by Dai et al. 2020.

Bias correction of low-frequency variability

Rocheta, Evans, and Sharma 2017 introduced the correction of low-frequency variability bias. This is especially helpful in the forecast of high-impact events, such as floods and droughts, which are usually related to low-frequency variability. It operates by replacing the lag-1 autocorrelation with the observed monthly lag-1 one. In the climatological study in which it was used, it showed the ability to improve the result but not significantly better than a simpler MVA (Xu, Y. Han, and Z. Yang 2019). Additional studies are needed to understand whether the inter-variable dependencies are maintained.

Spectral nudging during integration

One further bias correction technique introduces spectral nudging during RCM integration to continuously apply a correction to it (Xu and Zong-Liang Yang 2015). The main advantage is the continued bias correction, which is introduced everywhere and not only as a boundary and initial condition. It uses a 4Dvar data assimilation technique to constantly force the RCM towards a corrected GCM.

The drawback of this method is the strong disturbance it can generate in the RCM, especially if not properly calibrated. This may affect inter-variable dependencies, thus introducing new biases, as explained in Xu, Y. Han, and Z. Yang 2019.

B.3 Ensemble recalibration techniques

Ensemble Recalibration techniques (RC) are a relatively simple way of assessing the representativeness of the observation and its predictability by a model ensemble.

One example is the Climate Conserving Recalibration (CCR), first introduced by Francisco J. Doblas-Reyes, Hagedorn, and Palmer 2005. The name is a reference to its ability to not introduce any systematic bias in both the mean and the variance of the model climatology (Weigel, Liniger, and Appenzeller 2009). The expression is shown in Equation B.3. Each observation is assumed to be the sum of a predictable signal and a Gaussian-distributed stochastic noise. This way a hypothetical distribution is built and when considering the ensemble of forecasts it should be statistically indistinguishable from the observation. In an ideal case, the distributions should correspond. However, this is not the case in a real scenario, where the ensemble spread does not necessarily correspond at all times to the spread associated with the noise in the observation. Supposing that is the case, the forecasts are said to be unreliable, and to correct the mean and spread of the ensemble they are rescaled appropriately.

$$x'_m = \rho\frac{\sigma_o}{\sigma_{\langle x_m\rangle}}\langle x_m\rangle + \sqrt{1-\rho^2}\frac{\sigma_o}{\sigma_m}(x_m - \overline{\langle x_m\rangle}) + \overline{\langle x_o\rangle} \tag{B.3}$$

The Ratio of Predictable Components (RPC) introduced by Eade et al. 2014 and shown in Equation B.4 is another example of RC. The predictable component is defined as the square root of the fraction of the predictable variance. The idea is to compare this quantity for the forecast with that of the observation. The latter can be estimated through the computation of the Pearson correlation between observations and the ensemble mean, which represents the predictor. The former can instead be estimated using the ratio of the average ensemble variance and the average one for a single member. If the forecast were perfect in representing the actual predictability of the system, the RPC would be equal to one. This technique has been used both for seasonal (Manzanas, J. M. Gutiérrez, Bhend, Hemri, F. J. Doblas-Reyes, Torralba, et al. 2019) and in seasonal-to-decadal scales (Eade et al. 2014).

$$x'_m = \rho\frac{\sigma_o}{\sigma_{\langle x_{model}\rangle}}(\langle x_m\rangle - \overline{\langle x_m\rangle}) + \sqrt{1-\rho^2}\frac{\sigma_o}{\sqrt{\sigma^2_{x_m-\overline{\langle x_m\rangle}}}}(x_m - \langle x_m\rangle) + \overline{\langle x_o\rangle} \tag{B.4}$$

Yet another approach is the ensemble MOS Recalibration (MOS-RC), which can be implemented in different ways. One option consists in using a linear regression between the ensemble mean and the corresponding observation, as done in Marcos et al. 2018. The derived parameters are then used to rescale the forecast standardized anomalies. The generally good performance and relatively low computational cost of this approach have been stated in Manzanas, J. M. Gutiérrez, Bhend, Hemri, F. J. Doblas-Reyes, Penabad, et al. 2020. A more resource-intensive alternative is the use of a non-homogeneous Gaussian regression, first introduced by Gneiting et al. 2005 and applied to seasonal forecasts by Tippett and Barnston 2008. The expression is given by Equation B.5, where α, β, γ and δ are parameters. They are obtained through the minimization of the ensemble CRPS, which is a commonly used quality metric, mentioned in Section 2.7.

$$x'_m = \alpha + \beta(\langle x_m\rangle - \overline{\langle x_m\rangle}) + \sqrt{\gamma^2 + \delta^2\sigma^2_{x_m}}(x_m - \langle x_m\rangle) \tag{B.5}$$

These methods have been compared by Manzanas, J. M. Gutiérrez, Bhend, Hemri, F. J. Doblas-Reyes, Torralba, et al. 2019 in which it was found that they perform similarly. Other more complex RC methods do exist but were not considered as they are thought to lead to overfitting. RC can significantly and effectively reduce the bias in a way comparable to simple BA methods, while also enhancing forecast reliability. The performance is also favourable when compared to more complex PP or MOS statistical downscaling techniques, as shown in Manzanas, J. M. Gutiérrez, Bhend, Hemri, F. J. Doblas-Reyes, Penabad, et al. 2020. One limit they present is their inapplicability to daily data, whereas that is possible, in general, for BA methods.

Appendix C

Score tables

This Appendix contains the score tables referred to in Chapter 3 and supports the discussion about the choice of the downscaling configuration. Other tables are omitted for brevity.

Simulation	**Domain**	**BOU**	**BOI**	**MEZ**	**SPT**
Base	d01	6.71	5.74	7.92	12.41
Coast	d01	6.76	5.82	7.93	12.24
Coast_PBL	d01	7.25	5.95	6.98	8.96
Coast_RAD	d01	7.01	6.03	8.16	12.08
Coast_VL	d01	6.97	5.92	7.49	11.40
URB	d01	7.16	6.08	7.75	12.11
URB_5-1	d01	5.09	5.96	8.03	13.53
URB_VL	d01	6.33	5.53	7.35	11.28
ERA5	d01	8.72	7.26	6.89	10.29
Base	d02	6.95	5.97	8.69	13.08
Coast	d02	6.94	5.87	8.67	12.94
Coast_PBL	d02	6.90	5.42	6.26	9.33
Coast_RAD	d02	7.68	6.94	9.43	14.02
Coast_VL	d02	5.56	4.39	7.64	11.66
URB	d02	7.34	6.45	8.22	12.94
URB_5-1	d02	5.66	6.83	8.66	14.79
URB_VL	d02	5.12	4.46	7.15	11.56
Base	d03	6.86	6.37	8.89	12.94
Coast	d03	7.00	6.33	8.68	12.64
Coast_PBL	d03	6.71	5.12	6.26	9.64
Coast_RAD	d03	7.56	7.14	10.10	13.75
Coast_VL	d03	5.68	4.30	7.25	11.78
URB	d03	7.50	7.00	8.25	13.08
URB_VL	d03	5.25	5.28	7.10	11.82

Table C.1: MAE of two metres relative humidity for the different simulations and domains, referred to the selected locations. This Table is referred to in Section 3.1.

Simulation	**Domain**	**BOU**	**BOI**	**MEZ**	**SPT**
Base	d01	0.96	-0.79	-3.94	-10.81
Coast	d01	0.99	-0.76	-3.79	-10.66
Coast_PBL	d01	5.41	3.65	1.01	-6.20
Coast_RAD	d01	0.77	-0.98	-3.45	-10.20
Coast_VL	d01	3.83	2.08	-1.48	-9.04
URB	d01	1.03	-0.72	-3.69	-10.47
URB_5-1	d01	-2.83	-4.58	-6.26	-12.96
URB_VL	d01	2.72	0.97	-1.51	-8.95
ERA5	d01	8.10	6.35	-3.51	-8.33
Base	d02	-0.38	-2.13	-6.30	-12.07
Coast	d02	-0.13	-1.88	-6.00	-11.75
Coast_PBL	d02	3.02	1.27	-1.62	-7.62
Coast_RAD	d02	-1.12	-2.88	-6.21	-11.88
Coast_VL	d02	3.12	1.37	-1.82	-9.14
URB	d02	-0.80	-2.55	-5.55	-11.41
URB_5-1	d02	-3.66	-6.00	-7.46	-14.41
URB_VL	d02	0.26	-1.49	-2.21	-9.40
Base	d03	-1.54	-3.41	-6.51	-11.71
Coast	d03	-1.37	-3.22	-6.21	-11.45
Coast_PBL	d03	2.17	0.37	-2.09	-8.15
Coast_RAD	d03	-2.22	-3.93	-6.37	-11.83
Coast_VL	d03	1.89	0.13	-2.29	-9.29
URB	d03	-2.00	-3.94	-5.68	-11.24
URB_VL	d03	-0.64	-3.14	-2.53	-9.53

Table C.2: BIAS of two metres relative humidity for the different simulations and domains, referred to the selected locations. This Table is referred to in Section 3.1.

Simulation	**Domain**	**BOI**	**MEZ**	**STG**	**SPT**
Base	d01	1.14	1.61	1.64	1.70
Coast	d01	1.16	1.64	1.64	1.72
Coast_PBL	d01	1.12	1.46	1.58	1.49
Coast_RAD	d01	1.35	2.05	1.90	2.16
Coast_VL	d01	0.99	1.42	1.54	1.55
URB	d01	1.27	1.60	1.64	1.69
URB_5-1	d01	2.14	1.81	1.46	1.82
URB_VL	d01	1.12	1.42	1.53	1.54
ERA5	d01	1.81	1.61	1.68	2.48
Base	d02	1.23	1.71	1.79	1.82
Coast	d02	1.22	1.71	1.79	1.83
Coast_PBL	d02	1.14	1.49	1.55	1.55
Coast_RAD	d02	1.48	2.20	2.06	2.31
Coast_VL	d02	0.98	1.55	1.61	1.64
URB	d02	1.36	1.71	1.79	1.84
URB_5-1	d02	2.58	1.97	1.54	1.98
URB_VL	d02	1.33	1.50	1.59	1.60
Base	d03	1.34	1.71	1.71	1.71
Coast	d03	1.33	1.69	1.70	1.72
Coast_PBL	d03	1.12	1.51	1.59	1.53
Coast_RAD	d03	1.66	2.22	1.90	2.25
Coast_VL	d03	1.08	1.57	1.64	1.64
URB	d03	1.67	1.71	1.71	1.80
URB_VL	d03	1.94	1.57	1.65	1.63

Table C.3: Mean Absolute Error of two metres temperature for the different simulations and domains, referred to the selected locations.

Simulation	Domain	BOI	MEZ	STG	SPT
Coast_VL	d01	-0.41	1.01	0.28	1.31
Coast_VL_Bis	d01	-0.28	0.97	0.36	1.24
ERA5	d01	-1.31	1.48	1.62	2.46
Coast_VL	d02	-0.43	1.04	0.72	1.33
Coast_VL_Bis	d02	-0.31	1.03	0.79	1.30
Coast_VL	d03	0.15	1.10	0.58	1.33
Coast_VL_Bis	d03	0.24	1.13	0.63	1.30

Table C.4: BIAS of two metres temperature for the two alternative configurations of the vertical levels. Every available location is considered. This Table is referred to in Section 3.1.

Simulation	Domain	BOI	MEZ	STG	SPT
Coast_VL	d01	1.45	1.88	1.93	1.97
Coast_VL_Bis	d01	1.32	1.84	1.88	1.92
ERA5	d01	2.18	1.92	2.13	2.73
Coast_VL	d02	1.32	1.97	2.05	2.05
Coast_VL_Bis	d02	1.28	1.90	2.00	2.00
Coast_VL	d03	1.36	1.99	2.07	2.08
Coast_VL_Bis	d03	1.39	1.92	1.98	2.04

Table C.5: RMSE of two metres temperature for the two alternative configurations of the vertical levels, referred to the selected locations. This Table is referred to in Section 3.1.

Simulation	**Domain**	**BOI**	**MEZ**	**STG**	**SPT**
URB	d01	-0.27	1.46	0.15	1.38
URB_larger	d01	-0.35	1.33	0.10	1.18
ERA5	d01	-1.01	1.63	1.12	2.36
URB	d02	-0.12	1.65	0.57	1.46
URB_larger	d02	-0.40	1.22	0.50	1.21
URB	d03	0.65	1.58	0.35	1.43
URB_larger	d03	0.38	1.55	0.21	1.20

Table C.6: BIAS for the simulation introducing the specific urban parametrization and its variation with larger domains, described in Section 2.4, referred to the selected locations. This Table is referred to in Section 3.1.

Simulation	**Domain**	**BOI**	**MEZ**	**STG**	**SPT**
URB	d01	1.22	1.54	1.74	1.71
URB_larger	d01	1.28	1.64	1.85	1.74
ERA5	d01	1.86	1.66	1.96	2.86
URB	d02	1.51	1.76	2.04	1.86
URB_larger	d02	1.48	1.80	1.96	1.86
URB	d03	2.11	1.75	1.84	1.77
URB_larger	d03	2.07	1.78	1.63	1.70

Table C.7: RMSE for the simulation introducing the specific urban parametrization and its variation with larger domains, described in Section 2.4, referred to the selected locations. This Table is referred to in Section 3.1.

Bibliography

AghaKouchak, A. et al. (Dec. 12, 2022). "Status and prospects for drought forecasting: opportunities in artificial intelligence and hybrid physical–statistical forecasting". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 380.2238, p. 20210288. ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.2021.0288.

Albers, John R. and Matthew Newman (Nov. 2019). "A Priori Identification of Skillful Extratropical Subseasonal Forecasts". en. In: *Geophysical Research Letters* 46.21, pp. 12527–12536. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2019GL085270.

Anderson, D. et al. (2007). "Development of the ECMWF seasonal forecast System 3". In: DOI: 10.21957/TZ11XYPF4.

Bougeault, P. and P. Lacarrere (Aug. 1989). "Parameterization of Orography-Induced Turbulence in a Mesobeta–Scale Model". en. In: *Monthly Weather Review* 117.8, pp. 1872–1890. ISSN: 0027-0644, 1520-0493. DOI: 10.1175/1520-0493(1989)117⟨1872:POOITI⟩2.0.CO;2.

Breeden, Melissa L. et al. (Oct. 2022). "The Spring Minimum in Subseasonal 2-m Temperature Forecast Skill over North America". In: *Monthly Weather Review* 150.10, pp. 2617–2628. ISSN: 0027-0644, 1520-0493. DOI: 10.1175/MWR-D-22-0062.1.

Brier, Glenn W. (Jan. 1950). "VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY". en. In: *Monthly Weather Review* 78.1, pp. 1–3. ISSN: 0027-0644, 1520-0493. DOI: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Bröcker, Jochen (July 2009). "Reliability, sufficiency, and the decomposition of proper scores". en. In: *Quarterly Journal of the Royal Meteorological Society* 135.643, pp. 1512–1519. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.456.

Brogno, Luigi et al. (May 15, 2023). *A Novel Framework for the Assessment of Heat-Wave Risks and Nature-Based Solutions (NBS) Impacts.* DOI: 10.5194/egusphere-egu23-7995.

Brunner, Lukas et al. (June 2018). "Dependence of Present and Future European Temperature Extremes on the Location of Atmospheric Blocking". en. In: *Geophysical Research Letters* 45.12, pp. 6311–6320. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2018GL077837.

C3S (2018). *ERA5 hourly data on single levels from 1940 to present.* DOI: 10.24381/CDS.ADBB2D47.

Campbell, Sharon et al. (Sept. 2018). "Heatwave and health impact research: A global review". en. In: *Health & Place* 53, pp. 210–218. ISSN: 13538292. DOI: 10.1016/j.healthplace.2018.08.017.

Cárdenas Belleza, Gabriel A, Marc F P Bierkens, and Michelle T H Van Vliet (Oct. 2023). "Sectoral water use responses to droughts and heatwaves: analyses from local to global scales for 1990–2019". In: *Environmental Research Letters* 18.10, p. 104008. ISSN: 1748-9326. DOI: 10.1088/1748-9326/acf82e.

Colette, A., R. Vautard, and M. Vrac (July 2012). "Regional climate downscaling with prior statistical correction of the global climate forcing". en. In: *Geophysical Research Letters* 39.13, 2012GL052258. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2012GL052258.

Copernicus Climate Change Service (2018). *Seasonal forecast daily and subdaily data on single levels.* DOI: 10.24381/CDS.181D637E.

Cornes, Richard C. et al. (Sept. 2018). “An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets”. en. In: *Journal of Geophysical Research: Atmospheres* 123.17, pp. 9391–9409. ISSN: 2169-897X, 2169-8996. DOI: 10.1029/2017JD028200.

Crochemore, Louise, Maria-Helena Ramos, and Florian Pappenberger (Sept. 2016). “Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts”. en. In: *Hydrology and Earth System Sciences* 20.9, pp. 3601–3618. ISSN: 1607-7938. DOI: 10.5194/hess-20-3601-2016.

Dai, Aiguo et al. (July 2020). “A new approach to construct representative future forcing data for dynamic downscaling”. en. In: *Climate Dynamics* 55.1-2, pp. 315–323. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-017-3708-8.

De Andrade, Felipe M., Caio A. S. Coelho, and Iracema F. A. Cavalcanti (May 2019). “Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models”. en. In: *Climate Dynamics* 52.9-10, pp. 5451–5475. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-018-4457-z.

Deilami, Kaveh, Md. Kamruzzaman, and Yan Liu (May 2018). “Urban heat island effect: A systematic review of spatio-temporal factors, data, methods, and mitigation measures”. en. In: *International Journal of Applied Earth Observation and Geoinformation* 67, pp. 30–42. ISSN: 15698432. DOI: 10.1016/j.jag.2017.12.009.

Diro, G. T., A. M. Tompkins, and X. Bi (Aug. 2012). “Dynamical downscaling of ECMWF Ensemble seasonal forecasts over East Africa with RegCM3”. en. In: *Journal of Geophysical Research: Atmospheres* 117.D16, 2011JD016997. ISSN: 0148-0227. DOI: 10.1029/2011JD016997.

Diro, Gulilat Tefera (Feb. 2016). “Skill and economic benefits of dynamical downscaling of ECMWF ENSEMBLE seasonal forecast over southern Africa with RegCM4”. en. In: *International Journal of Climatology* 36.2, pp. 675–688. ISSN: 0899-8418, 1097-0088. DOI: 10.1002/joc.4375.

Doblas-Reyes, Francisco J., Renate Hagedorn, and T. N. Palmer (May 2005). “The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination”. en. In: *Tellus A* 57.3, pp. 234–252. ISSN: 0280-6495, 1600-0870. DOI: 10.1111/j.1600-0870.2005.00104.x.

Doblas-Reyes, F. J. et al. (July 2009). “Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts”. en. In: *Quarterly Journal of the Royal Meteorological Society* 135.643, pp. 1538–1559. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.464.

Dobrynin, Mikhail et al. (Apr. 2018). “Improved Teleconnection-Based Dynamical Seasonal Predictions of Boreal Winter”. en. In: *Geophysical Research Letters* 45.8, pp. 3605–3614. ISSN: 0094-8276, 1944-8007. DOI: 10.1002/2018GL077209.

Domeisen, Daniela I. V. et al. (Dec. 2022). “Prediction and projection of heatwaves”. en. In: *Nature Reviews Earth & Environment* 4.1, pp. 36–50. ISSN: 2662-138X. DOI: 10.1038/s43017-022-00371-z.

Eade, Rosie et al. (Aug. 2014). “Do seasonal-to-decadal climate predictions underestimate the predictability of the real world?” en. In: *Geophysical Research Letters* 41.15, pp. 5620–5628. ISSN: 0094-8276, 1944-8007. DOI: 10.1002/2014GL061146.

ECMWF (2016). “IFS Documentation CY41R2 - Part V: Ensemble Prediction System”. In: Publisher: ECMWF. DOI: 10.21957/4BTQAUG2X.

Epstein, Edward S. (Jan. 1969). “Stochastic dynamic prediction^1”. In: *Tellus A: Dynamic Meteorology and Oceanography* 21.6, p. 739. ISSN: 1600-0870. DOI: 10.3402/tellusa.v21i6.10143.

Ferranti, Laura and Pedro Viterbo (Aug. 2006). “The European Summer of 2003: Sensitivity to Soil Water Initial Conditions”. en. In: *Journal of Climate* 19.15, pp. 3659–3680. ISSN: 1520-0442, 0894-8755. DOI: 10.1175/JCLI3810.1.

Fischer, E. M. et al. (Mar. 2007). “Contribution of land-atmosphere coupling to recent European summer heat waves”. en. In: *Geophysical Research Letters* 34.6, 2006GL029068. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2006GL029068.

Folland, Chris K. et al. (Mar. 2009). “The Summer North Atlantic Oscillation: Past, Present, and Future”. en. In: *Journal of Climate* 22.5, pp. 1082–1103. ISSN: 1520-0442, 0894-8755. DOI: 10.1175/2008JCLI2459.1.

Fragkoulidis, G. et al. (Jan. 2018). “Linking Northern Hemisphere temperature extremes to Rossby wave packets”. en. In: *Quarterly Journal of the Royal Meteorological Society* 144.711, pp. 553–566. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.3228.

Francis, R. I. C. C. and J. A. Renwick (Sept. 3, 1998). “A Regression-based Assessment of the Predictability of New Zealand Climate Anomalies”. In: *Theoretical and Applied Climatology* 60.1, pp. 21–36. ISSN: 0177-798X, 1434-4483. DOI: 10.1007/s007040050031.

Ghimire, Uttam, Govindarajalu Srinivasan, and Anshul Agarwal (Mar. 2019). “Assessment of rainfall bias correction techniques for improved hydrological simulation”. en. In: *International Journal of Climatology* 39.4, pp. 2386–2399. ISSN: 0899-8418, 1097-0088. DOI: 10.1002/joc.5959.

Gneiting, Tilmann et al. (May 2005). “Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation”. en. In: *Monthly Weather Review* 133.5, pp. 1098–1118. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/MWR2904.1.

Golian, Saeed and Conor Murphy (Aug. 2022). “Evaluating Bias-Correction Methods for Seasonal Dynamical Precipitation Forecasts”. In: *Journal of Hydrometeorology* 23.8, pp. 1350–1363. ISSN: 1525-755X, 1525-7541. DOI: 10.1175/JHM-D-22-0049.1.

Han, Ji-Young et al. (Aug. 2023). “Ensemble size versus bias correction effects in subseasonal-to-seasonal (S2S) forecasts”. en. In: *Geoscience Letters* 10.1, p. 37. ISSN: 2196-4092. DOI: 10.1186/s40562-023-00292-9.

Hernández-Díaz, Leticia et al. (Apr. 2017). “3-Step dynamical downscaling with empirical correction of sea-surface conditions: application to a CORDEX Africa simulation”. en. In: *Climate Dynamics* 48.7-8, pp. 2215–2233. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-016-3201-9.

Hersbach, Hans et al. (July 2020). “The ERA5 global reanalysis”. en. In: *Quarterly Journal of the Royal Meteorological Society* 146.730, pp. 1999–2049. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.3803.

Hoffmann, P. et al. (Nov. 2016). “Bias and variance correction of sea surface temperatures used for dynamical downscaling”. en. In: *Journal of Geophysical Research: Atmospheres* 121.21. ISSN: 2169-897X, 2169-8996. DOI: 10.1002/2016JD025383.

Holland, Greg et al. (May 2010). “Model Investigations of the Effects of Climate Variability and Change on Future Gulf of Mexico Tropical Cyclone Activity”. In: *All Days*. Houston, Texas, USA: OTC, OTC–20690–MS. DOI: 10.4043/20690-MS.

Johnson, Stephanie J. et al. (Mar. 2019). “SEAS5: the new ECMWF seasonal forecast system”. en. In: *Geoscientific Model Development* 12.3, pp. 1087–1117. ISSN: 1991-9603. DOI: 10.5194/gmd-12-1087-2019.

Kain, John S. (Jan. 2004). “The Kain–Fritsch Convective Parameterization: An Update”. en. In: *Journal of Applied Meteorology* 43.1, pp. 170–181. ISSN: 0894-8763, 1520-0450. DOI: 10.1175/1520-0450(2004)043⟨0170:TKCPAU⟩2.0.CO;2.

Kautz, Lisa-Ann et al. (Mar. 2022). “Atmospheric blocking and weather extremes over the Euro-Atlantic sector – a review”. en. In: *Weather and Climate Dynamics* 3.1, pp. 305–336. ISSN: 2698-4016. DOI: 10.5194/wcd-3-305-2022.

Kim, Hyemi, Frédéric Vitart, and Duane E. Waliser (Dec. 2018). “Prediction of the Madden–Julian Oscillation: A Review”. en. In: *Journal of Climate* 31.23, pp. 9425–9443. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-18-0210.1.

Koster, R. D., S. P. P. Mahanama, T. J. Yamada, Gianpaolo Balsamo, A. A. Berg, M. Boisserie, P. A. Dirmeyer, F. J. Doblas-Reyes, et al. (Oct. 2011). “The Second Phase of the Global Land–Atmosphere Coupling Experiment: Soil Moisture Contributions to Subseasonal Forecast Skill”. en. In: *Journal of Hydrometeorology* 12.5, pp. 805–822. ISSN: 1525-755X, 1525-7541. DOI: 10.1175/2011JHM1365.1.

Koster, R. D., S. P. P. Mahanama, T. J. Yamada, Gianpaolo Balsamo, A. A. Berg, M. Boisserie, P. A. Dirmeyer, F. J. Doblas-Reyes, et al. (Jan. 2010). "Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment". en. In: *Geophysical Research Letters* 37.2, 2009GL041677. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2009GL041677.

Kowal, Katherine M. et al. (Jan. 2024). "Process-Informed Subsampling Improves Subseasonal Rainfall Forecasts in Central America". en. In: *Geophysical Research Letters* 51.1, e2023GL105891. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2023GL105891.

Kumar, Arun, Anthony G. Barnston, and Martin P. Hoerling (Apr. 2001). "Seasonal Predictions, Probabilistic Verifications, and Ensemble Size". en. In: *Journal of Climate* 14.7, pp. 1671–1676. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/1520-0442(2001)014⟨1671:SPPVAE⟩2.0.CO;2.

Lauwaet, D. et al. (Apr. 2012). "The precipitation response to the desiccation of Lake Chad". en. In: *Quarterly Journal of the Royal Meteorological Society* 138.664, pp. 707–719. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.942.

Lavaysse, Christophe et al. (Feb. 2019). "Predictability of the European heat and cold waves". en. In: *Climate Dynamics* 52.3-4, pp. 2481–2495. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-018-4273-5.

Leander, Robert and T. Adri Buishand (Jan. 2007). "Resampling of regional climate model output for the simulation of extreme river flows". en. In: *Journal of Hydrology* 332.3-4, pp. 487–496. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2006.08.006.

Lenderink, G., A. Buishand, and W. Van Deursen (May 2007). "Estimates of future discharges of the river Rhine using two scenario methodologies: direct versus delta approach". en. In: *Hydrology and Earth System Sciences* 11.3, pp. 1145–1159. ISSN: 1607-7938. DOI: 10.5194/hess-11-1145-2007.

Leung, L. Ruby et al. (Nov. 1999). "Simulations of the ENSO Hydroclimate Signals in the Pacific Northwest Columbia River Basin". en. In: *Bulletin of the American Meteorological Society* 80.11, pp. 2313–2329. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/1520-0477(1999)080⟨2313:SOTEHS⟩2.0.CO;2.

Libonati, Renata et al. (Jan. 2022). "Assessing the role of compound drought and heatwave events on unprecedented 2020 wildfires in the Pantanal". In: *Environmental Research Letters* 17.1, p. 015005. ISSN: 1748-9326. DOI: 10.1088/1748-9326/ac462e.

López-Espinoza, Erika Danaé et al. (Nov. 18, 2020). "Assessing the Impact of Land Use and Land Cover Data Representation on Weather Forecast Quality: A Case Study in Central Mexico". In: *Atmosphere* 11.11, p. 1242. ISSN: 2073-4433. DOI: 10.3390/atmos11111242.

Lupo, Anthony R. (Nov. 2021). "Atmospheric blocking events: a review". en. In: *Annals of the New York Academy of Sciences* 1504.1, pp. 5–24. ISSN: 0077-8923, 1749-6632. DOI: 10.1111/nyas.14557.

Madden, Roland A. (July 1976). "Estimates of the Natural Variability of Time-Averaged Sea-Level Pressure". en. In: *Monthly Weather Review* 104.7, pp. 942–952. ISSN: 0027-0644, 1520-0493. DOI: 10.1175/1520-0493(1976)104⟨0942:EOTNVO⟩2.0.CO;2.

Manual on the WMO Integrated Processing and Prediction System (2023). eng. OCLC: 741513974. Geneva: World Meteorological Organization.

Manzanas, R., M. D. Frías, et al. (Feb. 2014). "Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill". en. In: *Journal of Geophysical Research: Atmospheres* 119.4, pp. 1708–1719. ISSN: 2169897X. DOI: 10.1002/2013JD020680.

Manzanas, R., J. M. Gutiérrez, J. Bhend, S. Hemri, F. J. Doblas-Reyes, E. Penabad, et al. (Mar. 2020). "Statistical adjustment, calibration and downscaling of seasonal forecasts: a case-study for Southeast Asia". en. In: *Climate Dynamics* 54.5-6, pp. 2869–2882. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-020-05145-1.

Manzanas, R., J. M. Gutiérrez, J. Bhend, S. Hemri, F. J. Doblas-Reyes, V. Torralba, et al. (Aug. 2019). "Bias adjustment and ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the C3S dataset”. en. In: *Climate Dynamics* 53.3-4, pp. 1287–1305. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-019-04640-4.

Manzanas, R., J.M. Gutiérrez, et al. (Jan. 2018). “Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: Added value for user applications”. en. In: *Climate Services* 9, pp. 44–56. ISSN: 24058807. DOI: 10.1016/j.cliser.2017.06.004.

Manzanas, R., A. Lucero, et al. (Feb. 2018). “Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts?” en. In: *Climate Dynamics* 50.3-4, pp. 1161–1176. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-017-3668-z.

Manzanas, R., V. Torralba, et al. (Sept. 2022). “On the Reliability of Global Seasonal Forecasts: Sensitivity to Ensemble Size, Hindcast Length and Region Definition”. en. In: *Geophysical Research Letters* 49.17, e2021GL094662. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2021GL094662.

Maraun, Douglas (Dec. 2016). “Bias Correcting Climate Change Simulations - a Critical Review”. en. In: *Current Climate Change Reports* 2.4, pp. 211–220. ISSN: 2198-6061. DOI: 10.1007/s40641-016-0050-x.

Marcos, Raül et al. (Jan. 2018). “Use of bias correction techniques to improve seasonal forecasts for reservoirs — A case-study in northwestern Mediterranean”. en. In: *Science of The Total Environment* 610-611, pp. 64–74. ISSN: 00489697. DOI: 10.1016/j.scitotenv.2017.08.010.

Mariotti, Annarita et al. (May 2020). “Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond”. In: *Bulletin of the American Meteorological Society* 101.5, E608–E625. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/BAMS-D-18-0326.1.

Marzban, Caren, Scott Sandgathe, and Eugenia Kalnay (Feb. 2006). “MOS, Perfect Prog, and Reanalysis”. en. In: *Monthly Weather Review* 134.2, pp. 657–663. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/MWR3088.1.

Mason, I. (1982). “A model for assessment of weather forecasts”. en. In: *Aust. Meteorol. Mag.* 30, pp. 291–303.

Materia, Stefano et al. (Sept. 3, 2024). “Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives”. In: *WIREs Climate Change*, e914. ISSN: 1757-7780, 1757-7799. DOI: 10.1002/wcc.914.

May, Ryan M. et al. (2022). “MetPy: A Meteorological Python Library for Data Analysis and Visualization”. In: *Bulletin of the American Meteorological Society* 103.10, E2273–E2284. DOI: 10.1175/BAMS-D-21-0125.1.

Meehl, Gerald A. et al. (Apr. 2021). “Initialized Earth System prediction from subseasonal to decadal timescales”. en. In: *Nature Reviews Earth & Environment* 2.5, pp. 340–357. ISSN: 2662-138X. DOI: 10.1038/s43017-021-00155-x.

Mendez, Maikel et al. (Feb. 2020). “Performance Evaluation of Bias Correction Methods for Climate Change Monthly Precipitation Projections over Costa Rica”. en. In: *Water* 12.2, p. 482. ISSN: 2073-4441. DOI: 10.3390/w12020482.

Merryfield, William J. et al. (June 2020). “Current and Emerging Developments in Subseasonal to Decadal Prediction”. In: *Bulletin of the American Meteorological Society* 101.6, E869–E896. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/BAMS-D-19-0037.1.

Meyer, Jonathan D. D. and Jiming Jin (May 2016). “Bias correction of the CCSM4 for improved regional climate modeling of the North American monsoon”. en. In: *Climate Dynamics* 46.9-10, pp. 2961–2976. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-015-2744-5.

Miloshevich, George et al. (2022). *Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data.* Version Number: 2. DOI: 10.48550/ARXIV.2208.00971.

Miralles, Diego G. et al. (May 2014). “Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation”. en. In: *Nature Geoscience* 7.5, pp. 345–349. ISSN: 1752-0894, 1752-0908. DOI: 10.1038/ngeo2141.

Mlawer, Eli J. et al. (July 1997). “Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave”. en. In: *Journal of Geophysical Research: Atmospheres* 102.D14, pp. 16663–16682. ISSN: 0148-0227. DOI: 10.1029/97JD00237.

Mori, Paolo et al. (Jan. 2021). “Downscaling of seasonal ensemble forecasts to the convection-permitting scale over the Horn of Africa using the ¡span style=”font-variant:small-caps;”¿WRF¡/span¿ model”. en. In: *International Journal of Climatology* 41.S1. ISSN: 0899-8418, 1097-0088. DOI: 10.1002/joc.6809.

Murphy, J. M. (Jan. 1988). “The impact of ensemble forecasts on predictability”. en. In: *Quarterly Journal of the Royal Meteorological Society* 114.480, pp. 463–493. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.49711448010.

Pal, Sujan et al. (Mar. 2019). “Credibility of Convection-Permitting Modeling to Improve Seasonal Precipitation Forecasting in the Southwestern United States”. In: *Frontiers in Earth Science* 7, p. 11. ISSN: 2296-6463. DOI: 10.3389/feart.2019.00011.

Pappenberger, F. et al. (June 2009). “The Skill of Probabilistic Precipitation Forecasts under Observational Uncertainties within the Generalized Likelihood Uncertainty Estimation Framework for Hydrological Applications”. en. In: *Journal of Hydrometeorology* 10.3, pp. 807–819. ISSN: 1525-7541, 1525-755X. DOI: 10.1175/2008JHM956.1.

Park, Chang-Kyun and Jonghun Kam (Feb. 2023). “Sub-Seasonal Experiment (SubX) Model-based Assessment of the Prediction Skill of Recent Multi-Year South Korea Droughts”. en. In: *Asia-Pacific Journal of Atmospheric Sciences* 59.1, pp. 69–82. ISSN: 1976-7633, 1976-7951. DOI: 10.1007/s13143-022-00307-z.

Pegion, Kathy et al. (Oct. 2019). “The Subseasonal Experiment (SubX): A Multimodel Subseasonal Prediction Experiment”. In: *Bulletin of the American Meteorological Society* 100.10, pp. 2043–2060. ISSN: 0003-0007, 1520-0477. DOI: 10.1175/BAMS-D-18-0270.1.

Perkins, S. E. and L. V. Alexander (July 2013). “On the Measurement of Heat Waves”. en. In: *Journal of Climate* 26.13, pp. 4500–4517. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-12-00383.1.

Pfahl, S. and H. Wernli (June 2012). “Quantifying the relevance of atmospheric blocking for co-located temperature extremes in the Northern Hemisphere on (sub-)daily time scales”. en. In: *Geophysical Research Letters* 39.12, 2012GL052261. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2012GL052261.

Pielke, Roger A. and Robert L. Wilby (Jan. 2012). “Regional climate downscaling: What’s the point?” en. In: *Eos, Transactions American Geophysical Union* 93.5, pp. 52–53. ISSN: 0096-3941, 2324-9250. DOI: 10.1029/2012EO050008.

Possega, Marco et al. (Dec. 1, 2022). “Observational evidence of intensified nocturnal urban heat island during heatwaves in European cities”. In: *Environmental Research Letters* 17.12, p. 124013. ISSN: 1748-9326. DOI: 10.1088/1748-9326/aca3ba.

Prodhomme, Chloé et al. (Apr. 2022). “Seasonal prediction of European summer heatwaves”. In: *Climate Dynamics* 58.7, pp. 2149–2166. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-021-05828-3.

Pyrina, Maria and Daniela I. V. Domeisen (Jan. 2023). “Subseasonal predictability of onset, duration, and intensity of European heat extremes”. en. In: *Quarterly Journal of the Royal Meteorological Society* 149.750, pp. 84–101. ISSN: 0035-9009, 1477-870X. DOI: 10.1002/qj.4394.

Quesada, Benjamin et al. (Oct. 2012). “Asymmetric European summer heat predictability from wet and dry southern winters and springs”. en. In: *Nature Climate Change* 2.10, pp. 736–741. ISSN: 1758-678X, 1758-6798. DOI: 10.1038/nclimate1536.

Ribeiro, Andreia Filipa Silva et al. (Apr. 2020). *Risk of crop failure due to compound dry and hot extremes estimated with nested copulas.* DOI: 10.5194/bg-2020-116.

Robinson, Peter J. (Apr. 2001). “On the Definition of a Heat Wave”. en. In: *Journal of Applied Meteorology* 40.4, pp. 762–775. ISSN: 0894-8763, 1520-0450. DOI: 10.1175/1520-0450(2001)040⟨0762:OTDOAH⟩2.0.CO;2.

Rocheta, Eytan, Jason P. Evans, and Ashish Sharma (Dec. 2017). "Can Bias Correction of Regional Climate Model Lateral Boundary Conditions Improve Low-Frequency Rainfall Variability?" en. In: *Journal of Climate* 30.24, pp. 9785–9806. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-16-0654.1.

Röthlisberger, Matthias, Stephan Pfahl, and Olivia Martius (Oct. 2016). "Regional-scale jet waviness modulates the occurrence of midlatitude weather extremes". en. In: *Geophysical Research Letters* 43.20. ISSN: 0094-8276, 1944-8007. DOI: 10.1002/2016GL070944.

Rousi, Efi et al. (July 2022). "Accelerated western European heatwave trends linked to more-persistent double jets over Eurasia". en. In: *Nature Communications* 13.1, p. 3851. ISSN: 2041-1723. DOI: 10.1038/s41467-022-31432-y.

Russo, Simone et al. (Nov. 27, 2014). "Magnitude of extreme heat waves in present climate and their projection in a warming world". In: *Journal of Geophysical Research: Atmospheres* 119.22. ISSN: 2169-897X, 2169-8996. DOI: 10.1002/2014JD022098.

Schwitalla, Thomas et al. (Dec. 2008). "Systematic errors of QPF in low-mountain regions as revealed by MM5 simulations". en. In: *Meteorologische Zeitschrift* 17.6, pp. 903–919. ISSN: 0941-2948. DOI: 10.1127/0941-2948/2008/0338.

Seneviratne, Sonia I. et al. (May 2010). "Investigating soil moisture–climate interactions in a changing climate: A review". en. In: *Earth-Science Reviews* 99.3-4, pp. 125–161. ISSN: 00128252. DOI: 10.1016/j.earscirev.2010.02.004.

Shukla, J. (Oct. 1998). "Predictability in the Midst of Chaos: A Scientific Basis for Climate Forecasting". en. In: *Science* 282.5389, pp. 728–731. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.282.5389.728.

Skamarock, William C. et al. (Mar. 2019). *A Description of the Advanced Research WRF Model Version 4*. en. Tech. rep. [object Object]. DOI: 10.5065/1DFH-6P97.

Slater, Louise J. et al. (May 15, 2023). "Hybrid forecasting: blending climate predictions with AI models". In: *Hydrology and Earth System Sciences* 27.9, pp. 1865–1889. ISSN: 1607-7938. DOI: 10.5194/hess-27-1865-2023.

Sousa, Pedro M. et al. (Jan. 2018). "European temperature responses to blocking and ridge regional patterns". en. In: *Climate Dynamics* 50.1-2, pp. 457–477. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-017-3620-2.

Suárez, Danilo A (Dec. 2021). *Pywinter*. Version 2.0.6.

Teutschbein, Claudia and Jan Seibert (Aug. 2012). "Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods". en. In: *Journal of Hydrology* 456-457, pp. 12–29. ISSN: 00221694. DOI: 10.1016/j.jhydrol.2012.05.052.

Tippett, Michael K. (Sept. 2008). "Comments on "The Discrete Brier and Ranked Probability Skill Scores"". en. In: *Monthly Weather Review* 136.9, pp. 3629–3633. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/2008MWR2594.1.

Tippett, Michael K. and Anthony G. Barnston (Oct. 2008). "Skill of Multimodel ENSO Probability Forecasts". en. In: *Monthly Weather Review* 136.10, pp. 3933–3946. ISSN: 0027-0644, 1520-0493. DOI: 10.1175/2008MWR2431.1.

Van Den Hurk, Bart et al. (Jan. 2012). "Soil moisture effects on seasonal temperature and precipitation forecast scores in Europe". In: *Climate Dynamics* 38.1, pp. 349–362. ISSN: 0930-7575, 1432-0894. DOI: 10.1007/s00382-010-0956-2.

Wallace, John M. and Peter Victor Hobbs (2006). *Atmospheric science: an introductory survey*. 2nd ed. International geophysics series v. 92. OCLC: ocm62421169. Amsterdam ; Boston: Elsevier Academic Press. 483 pp. ISBN: 978-0-12-732951-2.

Weigel, Andreas P., Mark A. Liniger, and Christof Appenzeller (Apr. 2009). "Seasonal Ensemble Forecasts: Are Recalibrated Single Models Better than Multimodels?" en. In: *Monthly Weather Review* 137.4, pp. 1460–1479. ISSN: 1520-0493, 0027-0644. DOI: 10.1175/2008MWR2773.1.

Weisheimer, A. and T. N. Palmer (July 2014). "On the reliability of seasonal climate forecasts". en. In: *Journal of The Royal Society Interface* 11.96, p. 20131162. ISSN: 1742-5689, 1742-5662. DOI: 10.1098/rsif.2013.1162.

White, Christopher J. et al. (July 2017). "Potential applications of subseasonal-to-seasonal (¡span style="font-variant:small-caps;"¿S2S¡/span¿) predictions". en. In: *Meteorological Applications* 24.3, pp. 315–325. ISSN: 1350-4827, 1469-8080. DOI: 10.1002/met.1654.

White, R. H. and R. Toumi (June 2013). "The limitations of bias correcting regional climate model inputs". en. In: *Geophysical Research Letters* 40.12, pp. 2907–2912. ISSN: 0094-8276, 1944-8007. DOI: 10.1002/grl.50612.

Wilks, Daniel S (1995). *Statistical methods in the atmospheric sciences.* Academic, San Diego, California.

Wulff, C. Ole and Daniela I. V. Domeisen (Oct. 2019). "Higher Subseasonal Predictability of Extreme Hot European Summer Temperatures as Compared to Average Summers". en. In: *Geophysical Research Letters* 46.20, pp. 11520–11529. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2019GL084314.

Wulff, C. Ole, Richard J. Greatbatch, et al. (Nov. 2017). "Tropical Forcing of the Summer East Atlantic Pattern". en. In: *Geophysical Research Letters* 44.21. ISSN: 0094-8276, 1944-8007. DOI: 10.1002/2017GL075493.

Xu, Zhongfeng, Ying Han, and Zongliang Yang (Feb. 2019). "Dynamical downscaling of regional climate: A review of methods and limitations". en. In: *Science China Earth Sciences* 62.2, pp. 365–375. ISSN: 1674-7313, 1869-1897. DOI: 10.1007/s11430-018-9261-5.

Xu, Zhongfeng and Zong-Liang Yang (Sept. 2012). "An Improved Dynamical Downscaling Method with GCM Bias Corrections and Its Validation with 30 Years of Climate Simulations". en. In: *Journal of Climate* 25.18, pp. 6271–6286. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/JCLI-D-12-00005.1.

Xu, Zhongfeng and Zong-Liang Yang (Apr. 2015). "A new dynamical downscaling approach with GCM bias corrections and spectral nudging". en. In: *Journal of Geophysical Research: Atmospheres* 120.8, pp. 3063–3084. ISSN: 2169-897X, 2169-8996. DOI: 10.1002/2014JD022958.

Zampieri, Matteo et al. (Nov. 2016). "Global assessment of heat wave magnitudes from 1901 to 2010 and implications for the river discharge of the Alps". In: *Science of The Total Environment* 571, pp. 1330–1339. ISSN: 00489697. DOI: 10.1016/j.scitotenv.2016.07.008.

Zonato, A. et al. (Mar. 2020). "Evaluating the performance of a novel WUDAPT averaging technique to define urban morphology with mesoscale models". en. In: *Urban Climate* 31, p. 100584. ISSN: 22120955. DOI: 10.1016/j.uclim.2020.100584.

Zorita, Eduardo and Hans Von Storch (Aug. 1999). "The Analog Method as a Simple Statistical Downscaling Technique: Comparison with More Complicated Methods". en. In: *Journal of Climate* 12.8, pp. 2474–2489. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/1520-0442(1999)012⟨2474:TAMAAS⟩2.0.CO;2.