Seasonal forecasting of snow resources at Alpine sites

Silvia Terzago[1], Giulio Bongiovanni[1,2], and Jost von Hardenberg[3,1]

[1]Institute of Atmospheric Sciences and Climate, National Research Council, Torino, Italy
[2]Department of Civil, Environmental and Mechanical Engineering, University of Trento, Trento, Italy
[3]Department of Environment, Land and Infrastructure Engineering, Politecnico di Torino, Torino, Italy

Correspondence: Silvia Terzago (s.terzago@isac.cnr.it)

Abstract. Climate warming in mountain regions is resulting in glacier shrinking, seasonal snow cover reduction, changes in the amount and seasonality of meltwater runoff, with consequences on water availability. Droughts are expected to become more severe in the future with economical and environmental losses both locally and downstream. Effective adaptation strategies involve multiple time scales, and seasonal forecasts can help in the optimization of the available snow/water resources with lead times of several months. We developed a prototype to generate seasonal forecasts of snow depth and snow water equivalent with starting date November 1^{st} and lead times of 7 months, so up to May 31^{st} of the following year. The prototype has been co-designed with end users in the field of water management, hydropower production and of mountain ski tourism, meeting their needs in terms of indicators, time resolution of the forecasts, visualization of the forecast outputs. In this paper we present the modelling chain, based on the seasonal forecasts of ECMWF and Météo-France seasonal prediction systems, made available through the Copernicus Climate Change Service (C3S) Climate Data Store. Seasonal forecasts of precipitation, near-surface air temperature, radiative fluxes, wind and relative humidity are bias-corrected and downscaled to three sites in the Western Italian Alps, and finally used as input for the physically-based multi-layer snow model SNOWPACK. Precipitation is bias-corrected with a quantile mapping method using ERA5 reanalysis as a reference and then downscaled with the RainFARM stochastic procedure in order to allow an estimate of uncertainties due to the downscaling method. The impacts of precipitation bias-adjustment and downscaling on the forecast skill has been investigated.

The skill of the prototype in predicting the deviation of monthly snow depth with respect to the normal conditions from November to May in each season of the hindcast period 1995-2015 are demonstrated using both deterministic and probabilistic metrics. Forecast skills are determined with respect to a simple forecasting method based on the climatology, and station measurements are used as reference data. The prototype shows good skills at predicting the tercile category, i.e. snow depth below- and above-normal, in the winter (lead time 2-3-4 months) and spring (lead times 5-6-7 months) ahead: snow depth is predicted with higher accuracy (Brier Skill Score) and higher discrimination (Area Under the ROC Curve Skill Score) with respect to a simple forecasting method based on the climatology. Ensemble mean monthly snow depth forecasts are significantly correlated with observations not only at short lead time 1 and 2 months (November and December) but also at lead time 5 and 6 months (March and April) when employing the ECMWFS5 forcing. Moreover the prototype shows skill at predicting extremely dry seasons, i.e. seasons with snow depth below the 10^{th} percentile, while skills at predicting snow depth above the 90^{th} percentile are model-, station- and score-dependent. The bias-correction of precipitation forecasts is essential in case of large biases in the global seasonal forecast system (MFS6) to reconstruct a realistic snow depth climatology; however,

no remarkable differences are found among the skill scores when the precipitation input is bias-corrected, downscaled or bias-corrected and downscaled compared to the case in which raw data are employed, suggesting that skill scores are weakly sensitive to the treatment of the precipitation input.

1 Introduction

Mountain snowpack provides a natural reservoir which supplies water in the warm season for a variety of uses, such as hydropower production and irrigated agriculture in and downstream of mountain areas. However warming trends often amplified in mountain regions (Pepin et al., 2015; Palazzi et al., 2019) have resulted in glacier shrinking, seasonal snow cover reduction and changes in the amount and seasonality of runoff in snow dominated and glacier-fed river basins (Pörtner et al., 2019). Future cryosphere changes are projected to affect water resources and their uses (Pörtner et al., 2019). Current warm winter seasons may become normal at the end of the 21^{st} century, and there is indication for droughts to become more severe in the future (Haslinger et al., 2014; Stephan et al., 2021; Stahl et al., 2016). Effective adaptation strategies to address and reduce future water scarcity involve multiple time scales, from the seasonal scale, for the optimization of the available water resources with few months lead time, to climate scales, for the long-term planning of water storage infrastructures and the diversification of mountain tourism activities (Calì Quaglia et al., 2021). In these wide range of time scales, seasonal predictions have been considered with growing interest for their potential to provide early warning of extreme seasons, and to enable decision makers to take necessary actions to minimize negative impacts.

The ability of the current seasonal forecasts systems at predicting the main meteorological variables (air temperature and precipitation) is generally limited in the extra-tropics (Mishra et al., 2019) and this is reflected on poor streamflow prediction (Greuell et al., 2018; Arnal et al., 2018; Wanders et al., 2019; Santos et al., 2021). Some skill is found for the winter season streamflow prediction in about 40% of the European domain (Arnal et al., 2018), while contrasting results are found for high altitude catchments, where the discharge is mostly related to snow and ice melt. Some studies highlighted better skill than surrounding areas (Anghileri et al., 2016; Santos et al., 2021), while others found poor streamflow predictions due to the lack of snowpack predictability in the Alpine region (Wanders et al., 2019). One of the issues in mountain streamflow forecasting is the lack of reliable information to initialize physically based streamflow models, for example in terms of distribution of snow water equivalent (SWE) and of soil moisture, and this often results in limited forecasting skill (Li et al., 2019). In addition to initialization issues, ensemble streamflow predictions generally employ hydrological models in which the representation of snow processes is simplified and snow accumulation and melt are poorly captured (Wanders et al., 2019). These studies highlight the importance of a reliable representation of mountain snowpack for improving streamflow forecasts in mountain areas. An original approach to seasonal hydrological forecasting in mountain areas is to change the focus from the prediction of instantaneous hydrological fluxes (rainfall, streamflow) to that of slowly varying, and probably more predictable, hydrological quantities, such as the snow water equivalent (Förster et al., 2018). Snowpack is a natural "integrator" of the climatic conditions over multiple days/months, so even if daily temperature and precipitation forecasts do not match the corresponding observations, the differences may compensate over monthly/seasonal time scales and allow for reasonable monthly/seasonal snowpack

forecasts. Several economic activities recognized a value in seasonal forecasts of mountain snow accumulation, either per se or as an indicator of the meltwater available in the season ahead: i) public water managers, who can prepare strategies to mitigate the negative effects of extremely dry or extremely wet seasons, ii) hydropower companies involved in reservoir management, who use forecasts of the snowpack evolution to decide whether to release or save water in the reservoir; iii) mountain ski resorts managers, for which seasonal snowpack predictions are relevant to estimate the amount of artificial snow to be produced (Marke et al., 2015) and have high saving potential (Köberl et al., 2021).

The seasonal predictability of snow-related variables has so far been rarely studied. Kapnick et al. (2018) explored the potential of predicting the snowpack in March with 8 months lead time (starting date July 1^{st}) in the western US, using three atmosphere-ocean general circulation models (AOGCM) at different resolutions (200, 50 and 25 km). That study showed a good correlation to observations in most parts of the area, demonstrating the feasibility of such kinds of forecasts. In the Alpine region, Förster et al. (2018) tested a method to derive deterministic predictions of the sign of February SWE anomalies, i.e. SWE below- or above-average, over the Inn headwaters catchment. They set up a rather simple framework in which a distributed water balance model driven by seasonal forecasts of monthly air temperature and precipitation anomalies provides SWE anomaly forecasts over the basin. This forecasting method showed some skill in predicting the sign of the basin-average SWE anomaly and, more in general, it proved the higher robustness of SWE predictions compared to precipitation ones. However the deterministic approach adopted in that study does not allow to obtain a quantification of the uncertainty of the forecasts, and the only information on the sign of the SWE anomaly without an associated probability of occurrence is of limited usefulness in practical applications. In complex modeling chains the accuracy of the output variables is subject to multiple sources of uncertainty, which are present in the various components of the modelling chain: the global forecast system(s) employed; bias adjustment eventually applied to adjust systematic errors in the global models; downscaling techniques eventually applied to mitigate the mismatch between the scale of the forcing and the scale at which snow processes occurs; the process model employed, its setup and initialization. Each component of the chain should be evaluated to assess its relative contribution to the overall forecasting error, however this analysis is often overlooked or not adequately performed.

In this study we present a method to generate for the first time multi-system multi-member seasonal forecasts of mountain snow depth/water equivalent during the period from November to May of the following year, taking advantage of the state-of-the-art modelling techniques. We developed a prototype which uses seasonal forecasts of the main meteorological variables produced by seasonal forecast systems of the Copernicus Climate Change Service (C3S) to simulate the snowpack evolution at a given mountain site. Seasonal forecast system outputs at 1°x1° spatial resolution and daily or 6-hourly temporal resolution are bias-corrected and downscaled using different techniques depending on the variable type and characteristics (i.e. instantaneous or flux variable) to generate km-scale, hourly forcings. This fine scale hourly forcing is employed to drive the physical, multi-layer, 1-dimensional snow model SNOWPACK (Lehning et al., 2002) which proved to be one of the best performing models in a recent benchmark study (Terzago et al., 2020). The prototype is run at each location, and at each location it provides ensembles of snow depth seasonal forecasts at hourly time step, which are then aggregated to monthly or seasonal scale for the analysis.

The prototype is demonstrated at three selected sites in the Western Italian Alps, where snow seasonal forecasts can be exploited by stakeholders in the fields of hydropower energy production, water management and ski resort management. Ensemble seasonal forecasts are evaluated using both deterministic and probabilistic metrics (Wilks, 2011) to assess different forecast features (accuracy, discrimination and sharpness) at the monthly and seasonal scales. The skill of the forecast system is assessed compared to a reference forecast based on the past observations at in-situ stations. We also present an evaluation of the uncertainty associated with each step of the modelling chain, for example verifying the impact of using meteorological inputs from different seasonal forecast systems, and alternatively applying bias-adjustment or downscaling methods or a combination of both.

The paper is organized as follows. Section 2 describes the study area, the data used, the modelling chain and the forecasting skill assessment methods. Section 3 presents the results in terms of forecast skill of the prototype, and it is followed by a discussion and final conclusions in Sect. 4 and 5, respectively.

2 Methodology

The prototype has been co-designed with stakeholders, who provided guidance on the features required to make this climate service useful for applications. Although the purpose of the prototype is to respond to specific needs of the users, it has been developed to be general, flexible and applicable to any area of study for which seasonal snow forecasts are needed. In the following we present the motivations for the study, that closely determine the area of evaluation of the prototype, the datasets employed and a step-by step description of the methodology.

2.1 Motivation for the work, domain of study and in-situ data

The prototype has been conceived for applications in the Western Italian Alps, in three Valleys which are relevant for different stakeholders (Fig. 1a), i.e. i) the Orco Valley, hosting an artificial water reservoir serving a plant for hydropower production; ii) the Ala Valley, relevant for water supply to the Metropolitan City of Torino, 2.2 million inhabitants; and iii) the Upper Sesia Valley, which hosts one of the largest ski resorts in the Western Italy, at the foot of Monte Rosa. All stakeholders are interested in seasonal forecasts of snow abundance to plan in advance activities and investments for the season ahead. In particular they are interested in forecasting low snow seasons to limit snow/water shortage and economic losses. Each area of study hosts at least one station which provides nivo-meteorological data since the 1990s useful to evaluate model outputs. For each station, Table 1 reports the name, the geographical position, variables provided and start/end of the station activity. All stations are situated at elevations above 2000 m a.s.l. and snow cover is present for most of the year (Fig. 1b). At these altitudes a critical variable to measure is total precipitation, which is typically underestimated by standard (unheated) pluviometers. A quality check of the station data showed that increases in snow depth are often associated with daily total precipitation equal or close to zero. This suggests that standard pluviometers strongly underestimate solid precipitation, so total precipitation measurements are considered unreliable during the snow season and they have not been used in the analysis.

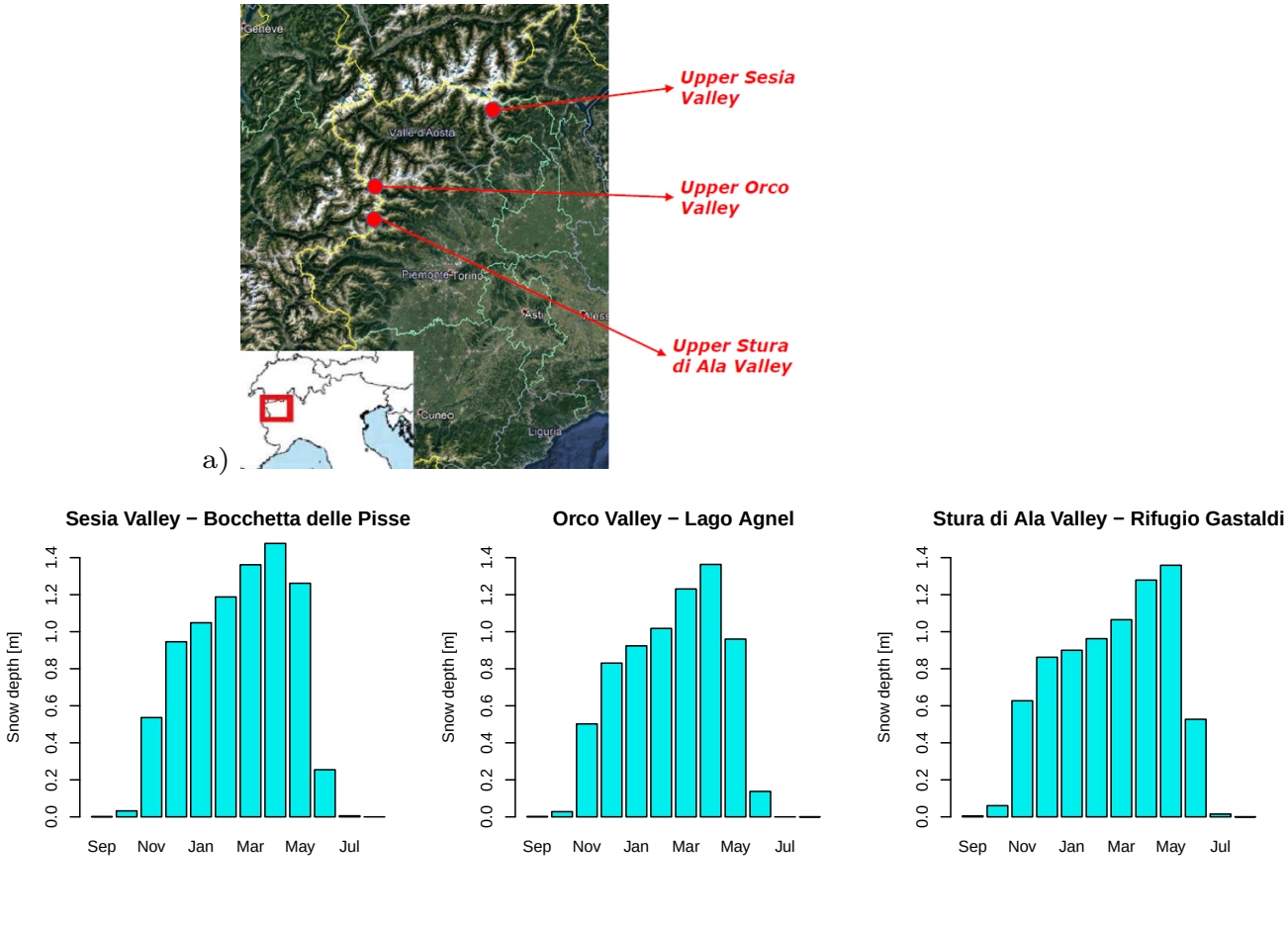

Figure 1. a) Map of the study sites indicating the three nivo-meteorological stations in NW Italian Alps (©Google Maps 2021). b) Snow depth climatology at the three stations considered in this study and described in Table 1. Averages are calculated over the period 1998-2015.

2.2 ERA5 reanalysis

In addition to observational data we use the latest ECMWF global reanalysis product, ERA5 (Hersbach et al., 2020), which provides reanalysis fields at 0.25° (about 30 km) spatial resolution and 1 hour temporal resolution. Compared to the previous reanalysis product, ERA-Interim, ERA5 uses one of the most recent versions of the Earth system model and data assimilation methods applied at ECMWF and modern parameterizations of Earth processes. With respect to ERA-Interim, ERA5 also has an improved global hydrological and mass balance, reduced biases in precipitation, and refinements of the variability and trends of surface air temperature (Hersbach et al., 2020). To supply the lack of continuous and/or trusted observational data, we use the ERA5 reanalysis at the gridpoint closest to each station to run reference simulations with the snow model. To this end, we perform two different ERA5-driven simulations differing by the air temperature input: in one case we use ERA5 raw temperature data, in the other case we use ERA5 bias-corrected temperature data, to which a simple mean bias correction with

Table 1. Stations considered in this study, elevation, position and date of start of automatic meteorological station records.

	Station		
	Bocchetta delle Pisse	**Lago Agnel**	**Rifugio Gastaldi**
Valley	Sesia	Orco	Stura di Ala
Elevation (m a.s.l.)	2410	2304	2659
Latitude (WGS 84, °)	45.875556	45.467778	45.298056
Longitude (WGS 84, °)	7.901111	7.139167	7.143333
Air temperature	01/01/1988	11/10/1996	30/04/1988
Total precipitation	06/07/1996	12/10/1996	05/07/1996
Wind Speed	01/01/1990	-	01/01/1990
Total incoming SW radiation	22/03/2012	-	06/10/2017
Snow depth	01/01/1995	01/11/1997	01/01/1995
Fresh snow depth	01/01/1995	01/11/1997	01/01/1995

respect to observations has been applied. In detail, the bias correction is carried out as follows: we derive the multi-annual average daily temperature bias of ERA5 with respect to observations, then we linearly interpolate the bias in time to the ERA5 resolution (1 hour) and we finally apply this offset to the original ERA5 hourly data. This simple method, hereafter referred to as the Mean Bias-Correction (MBC), allows to successfully reproduce (by construction) the observed temperature seasonal cycle (Fig. 2a) and it also implicitly takes into account scaling issues due to the different resolution of ERA5 and observational data. ERA5-driven snow depth simulations employing these two different temperature input data, together with snow depth measurements, are the benchmark against which we evaluate the seasonal snow depth forecasts.

2.3 Seasonal forecast data

We employ historical forecasts (hindcasts) from ECMWF System 5 (ECMWFS5, Johnson et al., 2019) and Météo-France System 6 (MFS6, Dorel et al., 2017) models obtained from the Copernicus Climate Data Store (https://climate.copernicus.eu/). For each system, we consider the 25-member hindcasts initialized each November 1^{st} and run for the 7 months ahead (November-May) over the period 1995-2015 (21 hindcasts) for which evaluation data (snow depth observations) were available from the stations. We consider all the variables needed to force the snow model: 2m temperature, 2m dewpoint temperature, total precipitation, surface solar and thermal radiation downwards, soil temperature level 1, 10-meter U and V wind components. Original C3S flux variables (precipitation and radiation) are accumulated since the beginning of the forecast, so they have been converted to daily values (see Table 2 for details). Horizontal wind components are converted to wind speed (modulus). Possible discrepancies between the climatologies of seasonal forecast and reference data (from observations, where available, or ERA5) have been investigated and adjusted using suitable methods as described in the following sections. Seasonal forecasts resolu-

Table 2. C3S seasonal forecast model variables used to create the forcing for the prototype: original variable name, short name and units, variable short name and units after post-processing (see Sect. 2.3 for details).

C3S variable	Short name	Units	Frequency	Short name CV*	Units CV*
2m temperature	t2m	K	6 h instantaneous	tas	K
2m dew point temperature	d2m	K	6 h instantaneous	tdps	K
Total precipitation	tp	m	24h aggregation**	prlr	mm/day
Surface solar radiation downwards	ssrd	J/m^2	24h aggregation**	rsds	W/m^2
Surface thermal radiation downwards	strd	J/m^2	24h aggregation**	rlds	W/m^2
Soil temperature level 1	tsl1	K	6 h instantaneous	tsl1	K
10 metre U and V wind components	u10, v10	m/s	6 h instantaneous	sfcWind	m/s

*CV=Converted variable

**=since beginning of forecast

tion is 1°Lon x 1°Lat in space and daily or 6 hourly in time. These resolutions are insufficient to simulate snow processes at the local scale, so we apply simple downscaling techniques to generate data at 1 km spatial resolution and 1 hour temporal resolution. The applied techniques are specific for each variable and they are briefly described in the following.

2.3.1 Air temperature

Figure 2a shows the multi-year average of the November-May 2m air temperature from ECMWFS5 hindcasts compared to observations. The ECMWFS5 temperature bias is large and time-dependent, and the same happens for MFS6 seasonal forecast system (not shown). To adjust the seasonal forecast system temperature bias we employ the mean bias-correction method used for ERA5 and based on the correction of the forecast data for the average daily bias with respect to observations (Sec. 2.2). The effect of the bias correction is displayed in Fig. 2a: the seasonal forecast system annual cycle appears very close to the observed one but it is smoother since it is averaged over all ensemble members. This simple approach has the advantage that it takes into account both the forecast system temperature bias and, implicitly, also scaling issues due to the different resolutions of model and observational data.

2.3.2 Total precipitation

Figure 2b shows the discrepancy between the ECMWFS5 daily precipitation climatologies and the ERA5 reference: the bias has been adjusted with a rather sophisticated approach which allows to take into account orographic effects. First, daily precipitation seasonal forecasts have been adjusted by applying quantile mapping (Gudmundsson et al., 2012; Perez-Zanon et al., 2021) on a monthly basis, using ERA5 total precipitation data upscaled to 1° as a reference dataset. Then bias-adjusted daily data have been downscaled from 1° to about 1 km using the RainFARM stochastic precipitation downscaling method (Rebora et al., 2006; D'Onofrio et al., 2014) improved to take into account orographic effects (Terzago et al., 2018). This method employs orographic weights derived from a fine-scale precipitation climatology (WorldClim, Fick and Hijmans, 2017) to correct

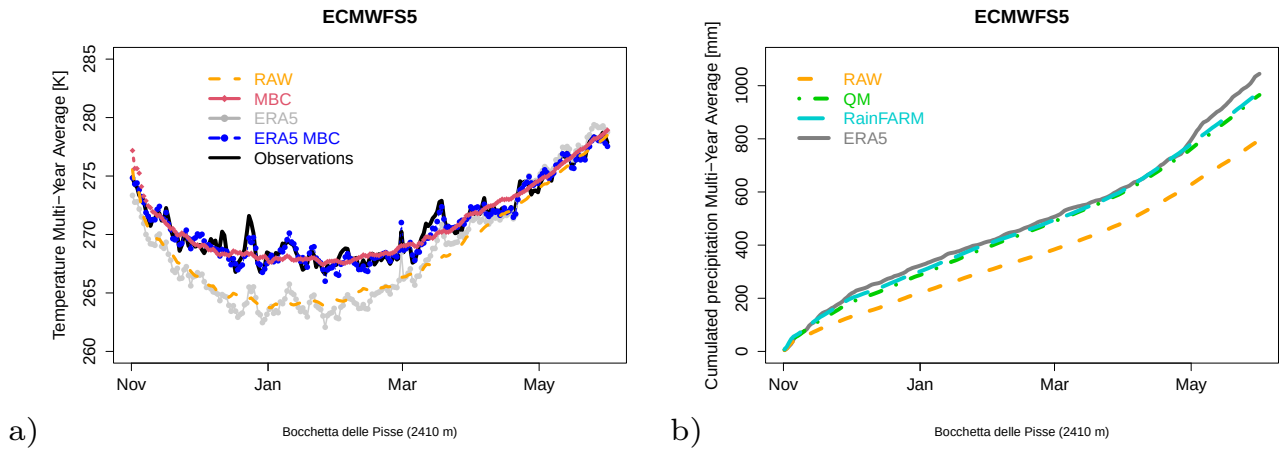

Figure 2. Multi-annual (1995-2015) averages at Bocchetta delle Pisse station (2410 m a.s.l.) of: a) daily air temperature in (gray) the ERA5 reanalysis, (blue) the ERA5 reanalysis bias-corrected with respect to observations with the delta method, (orange) ECMWFS5 seasonal forecasts, (red) ECMWFS5 seasonal forecast after the bias correction with respect to observation with the delta method, (black) observations; b) accumulated total precipitation in (gray) the ERA5 reanalysis and in ECMWFS5 seasonal forecasts with different levels of post-processing: (green) after the bias correction with the quantile mapping with respect to ERA5 method at the monthly scale, (cyan) after the bias correction and the downscaling with the RainFARM method adapted for complex terrains.

the downscaled field (Terzago et al., 2018). The RainFARM method is used to generate an ensemble of 10 stochastic realizations of the downscaled precipitation for each of the 25 seasonal forecast system ensemble members. This procedure allows generating 250-member ensemble forecasts for each starting date. Looking at the results in Fig. 2b, the quantile mapping allows to accurately reconstruct the long term climatology of the accumulated precipitation, and this feature is conserved after the application of the RainFARM downscaling. After the application of the spatial downscaling, precipitation is then disaggregated in time, from daily to hourly resolution, by equally redistributing the precipitation amount over all time steps with sufficient relative humidity to allow precipitation. We chose RH>80% as a threshold.

2.3.3 Surface shortwave and longwave radiation downwards

Daily accumulated surface shortwave and longwave radiation downwards (J/m^2) have been converted into average daily radiation fluxes (W/m^2) and downscaled in space using a simple bilinear interpolation to the coordinates of the station using the Climate Data Operator command line tools (CDO, Schulzweida, 2019). The effects of local terrain features such as the elevation difference between the model gridpoint and the station, the sky view factor and the terrain shading are not taken into account with this simple method, making the hypothesis that the uncertainty introduced by this simplification is much smaller than the uncertainty of the forecasts. In order to disaggregate in time average daily fluxes into hourly fluxes we employed a sort of analogue method using ERA5 as a reference. The choice of ERA5 as a reference dataset is supported by a high temporal correlation with both seasonal forecast systems, for each station and for both shortwave (r>0.82) and longwave (r>0.51) radia-

tion. For each day of the forecast period i) we consider the seasonal forecast of (shortwave, longwave) daily radiation for that day, ii) we consider all ERA5 daily average radiation values for that month over the period 1993-2019, iii) we sort the ERA5 daily values in ascending order, from the lowest to the highest, iii) we consider the 11 ERA5 values closest to the forecast for that day, iv) we randomly choose one among the 11 ERA5 daily values and we consider the corresponding 24 hourly values, v) we assume these 24 ERA5 hourly values to be the seasonal forecast of hourly radiation for that day. This technique allows us to reconstruct hourly forecasts which are plausible for the specific month and which conserve the daily mean radiation forecast for that specific day.

2.3.4 Humidity, surface wind and soil temperature

Seasonal forecast models in the CDS archive do not provide directly specific or relative humidity among their output variables. So we derive relative humidity from air temperature and dew point temperature following Lawrence (2005). Air temperature, dew point temperature as well as wind speed and soil temperature have been bilinearly interpolated to the coordinates of the station.

2.4 The SNOWPACK model

We simulate snow dynamics with the SNOWPACK model, a sophisticated snow and land-surface model allowing for a detailed description of the mass and energy exchange between the snow, the atmosphere and optionally the vegetation cover and the soil (Bartelt and Lehning, 2002). It provides a detailed description of snow properties, including weak layer characterization, phase changes and water transport in snow (Hirashima et al., 2010). A particular feature is the treatment of soil and snow as a continuum with a choice of a few up to several hundred layers (Bartelt and Lehning, 2002). The model is able to accurately estimate mountain snow depth in a variety of meteorological conditions, with an average error of about 10 cm when forced by accurate in-situ data (Terzago et al., 2020). The SNOWPACK model is used in its default configurations, so no tuning of the model parameters is made to improve the snow depth simulations locally. The snowpack lower boundary conditions are provided in terms of ground temperature in the topmost part of the soil at the soil-snow interface. We assume that the presence of a thick, insulating snowpack during the simulation period (Fig. 1b) decouples soil and atmospheric dynamics, thus ground and soil temperatures remain close to 0°C and deep soil layers do not affect the snowpack dynamics (Wever et al., 2015).

In our framework the SNOWPACK model has to be initialized with measured snow depth on November 1^{st}. In most seasons at the site considered the snow onset is in October, and on November 1^{st} the snowpack is already well established (i.e. snow depth $\geq$ 10 cm) as shown in Fig. 1b. In such cases SNOWPACK is initialized with the observed snow depth and a snow profile which characterizes each snow layer. Since the snow profile is unavailable from observations we simulate it by running SNOWPACK over the previous summer and driving the model with a mix of reanalysis and observational data: all meteorological forcing are provided by ERA5 except for air temperature which is derived by observations. Simulations generally start on August 1^{st}, or the following first day with observed snow depth SD=0, and end on 1^{st} November, providing the snow profile for that day, which is then used to initialize the SNOWPACK simulation in forecast mode. Otherwise, in the remaining seasons for which on November 1^{st} snow depth is lower than 10 cm, i.e. snow cover is shallow/discontinuous/absent, the SNOWPACK

Table 3. Plan of experiments with the SNOWPACK model. The meteorological forcing is generated using ECMWFS5 and MFS6 seasonal forecast systems outputs

Experiment	Total precipitation	Output ensemble members
RAW	RAW	25
QM	Quantile Mapping (reference ERA5)	25
RainFARM	RainFARM	250
QM+RainFARM	Quantile Mapping + RainFARM	250

model is initialized with snow depth equal to zero and run in forecast mode over the season ahead. Shallow snow cover has been aligned to snow free soil due to the difficulty of reliably simulating such thin snow covers.

2.5 Experiments with the SNOWPACK model

Precipitation is a critical parameter both to measure and to represent in model simulations. As explained in Sect. 2.3.2 we employ quite sophisticated techniques to bias-adjust and downscale precipitation forecasts to the station scale. Such complexity could be a limit in an operational framework where simple, easy-to-use approaches are preferred. To this aim we investigate a range of methods to correct precipitation inputs to verify if simpler methods can provide comparable results with respect to the most complex ones. We devised a set of 4 experiments with the SNOWPACK model, differing in the treatment of the precipitation input, with the aim of evaluating the model sensitivity to the accuracy of the precipitation input. The experiments are reported in Table 3 and briefly summarized here: 1) the first experiment (RAW) uses original seasonal forecast precipitation data without any further refinement; 2) the second experiment (QM) uses precipitation data bias-adjusted with the quantile mapping method using ERA5 as a reference dataset; 3) the third experiment (RainFARM) uses seasonal forecast precipitation data stochastically downscaled to 1 km with the RainFARM method; 4) the last experiment (QM+RainFARM) uses both the quantile mapping and the RainFARM methods to bias-adjust and downscale precipitation forecasts. For each experiment and each seasonal forecast system we run the modelling chain on a set of 21 meteorological forecasts starting on November 1^{st} of each year in the period 1995-2015.

2.6 Output of the modelling chain

For each experiment of Table 3, the output of the modelling chain consists of an ensemble of hourly (or eventually daily) snow depth time series representing the seasonal forecasts for the three considered stations. The number of ensemble members is 25 in the RAW and QM experiments and 250 in the RainFARM and QM+RainFARM experiments, i.e. 10 RainFARM precipitation downscaling realizations for each of the 25 model ensemble members (Table 3). An example of ensemble snow depth seasonal forecast for the season 2006/2007 is reported in Fig. 3 and it will be discussed in Sect. 3. In order to perform the statistical analysis of the set of snow depth hindcasts, the output of the modelling chain originally at hourly time step is

aggregated at the daily, monthly and seasonal (December to February (DJF), March to May (MAM) and November to May (NM)) scale to be compared with in-situ station measurements.

2.7 Evaluation metrics

Hourly snow depth seasonal forecasts are first aggregated to daily data and then to monthly and seasonal means over winter (DJF), spring (MAM) and the full November-May (NM) season. The seasonal means are computed by using all corresponding daily data. Monthly and seasonal forecasts are then evaluated by employing both deterministic and probabilistic metrics. While deterministic metrics consider the ensemble mean of the forecasts compared to the observations, probabilistic metrics compare different features of the forecast distribution with respect to the observations or the observed distribution. In the following we briefly describe all the metrics used in this study:

- Time correlation: The simplest way to evaluate ensemble forecasts is to assess the time correlation between ensemble mean forecasts and observations. Since we are interested in assessing the correlation of fluctuations, the linear trend in time series has been removed and the correlation has been calculated on residuals. The correlation is expressed as Pearson's correlation coefficient, the confidence interval is computed by a Fisher transformation and the significance level relies on a one-sided student-T distribution, with threshold 0.95 (BSC-CNS et al., 2021)

- Brier Score (BS): Among the set of probabilistic scores the Brier Score represents the mean square error of the probability forecast for a binary event, e.g. snow depth in a given tercile of the distribution (Mason, 2004). In our analysis, continuous forecasts are first transformed into tercile-based forecasts (i.e. probabilities for snow depth forecast to fall into the lower, middle or upper tercile of the forecast distribution) as suggested in Mason (2018). Then, the BS is calculated for each tercile. We also explored the forecast skill in predicting extreme events, i.e. the BS associated to monthly and seasonal snow depth below the 10^{th}- and above the 90^{th}-percentile of the forecast distribution. Tercile and percentile thresholds are calculated over the reference period 1995-2015

- Area Under the ROC curve (AUC): The Receiver Operating Characteristic Curve (ROC, Jolliffe and Stephenson, 2012) similarly to the Brier Score, allows the evaluation of binary forecasts. Given an ensemble forecast for a binary event, for example snow depth in the upper tercile, the ROC curve shows the true-positive rate against the false-positive rate for different probability threshold settings. The area under the ROC curve, shows the ability of the forecast system to discriminate between "event" and "non-event", i.e. it is a measure of the discrimination of the forecast system. AUCs are calculated separately for each tercile and then averaged over the three terciles

- Continuous Ranked Probability Score (CRPS): One of the most widely used accuracy metrics for ensemble forecasts is the Continuous Ranked Probability Score (Matheson and Winkler, 1976). The CRPS is the integrated squared difference between the forecast cumulative distribution function (CDF) and the empirical (observed) CDF, which is a step function. The CRPS has a negative orientation, i.e. the lower the score the better the forecast CDF approximates the observed CDF. The perfect value for CRPS is 0.

To facilitate the interpretation of the results of ensemble forecasts evaluation, the BS, AUC and CRPS scores are presented in terms of skill scores (SS). The skill scores indicate the skill of the forecast method with respect to a reference "trivial" forecast method based for example on the climatology, the persistence of the observed anomaly, etc. In our case the reference is the (monthly or seasonal) climatological forecast, derived from the set of climatological values except for the value that occurred. The sign and the absolute value of the skill score provide information on the added value of the forecast method compared to the climatological forecast: the more positive is the skill score, the better is the quality of the forecast; the more negative is the skill score the worse is the quality of the forecast; a skill score of 0 indicates no improvements with respect to the reference forecast; a skill score of 1 would instead indicate a perferct forecast. The analysis in terms of skill scores provides a quantitative and rigorous information on the quality and the different features of the forecast method. BSS and CRPS are calculated for each starting date and lead time, then averaged over all starting dates and converted into skill scores as follows:

$$SS = \frac{S - S_{ref}}{S_{perf} - S_{ref}} \tag{1}$$

where SS is the value of the skill score, S is the value of the score of the forecast system against the observations, S_{ref} is the value of the score of the climatological forecast against the observations and S_{perf} is the value of the score in the theoretical case that forecasts perfectly match observations. The AUC Skill Score (AUCSS), instead, is derived using the following formula (Wilks, 2011):

$$AUCSS = 2(AUC - 0.5) \tag{2}$$

The uncertainty on the time correlation and the skill scores has been evaluated by estimating the confidence interval (CI) using the bootstrap method (Bradley et al., 2008; Wilks, 2011), as recommended by Mason (2018). Bootstrapping is widely used to find the sampling distribution of a quantity and then to compute its standard error and CI. At first, given n the number of ensemble members, depending on whether n is odd or even, $n/2$ or $(n+1)/2$ members are randomly selected with replacement. Thus, a skill score is computed considering only selected ensemble members. The procedure has been iterated 1000 times generating a sample distribution, from which mean and 90% confidence interval error bars are estimated.

3 Results

3.1 An example of snow depth forecast

Figure 3 represents an example of snow depth forecast for the season 2006/2007 referring to the station of Bocchetta delle Pisse. The forecast is derived using the meteorological forcing provided by ECMWFS5, post-processed as described in Sect. 2.3. Precipitation forecasts have been bias-adjusted with the quantile mapping method and then downscaled to 1 km with the RainFARM method (QM+RainFARM experiment) generating 10 stochastic realizations for each of the 25 forecast ensemble member (250 downscaled precipitation forecasts in total). The ensemble spread, the 5-95th percentile range and the ensemble

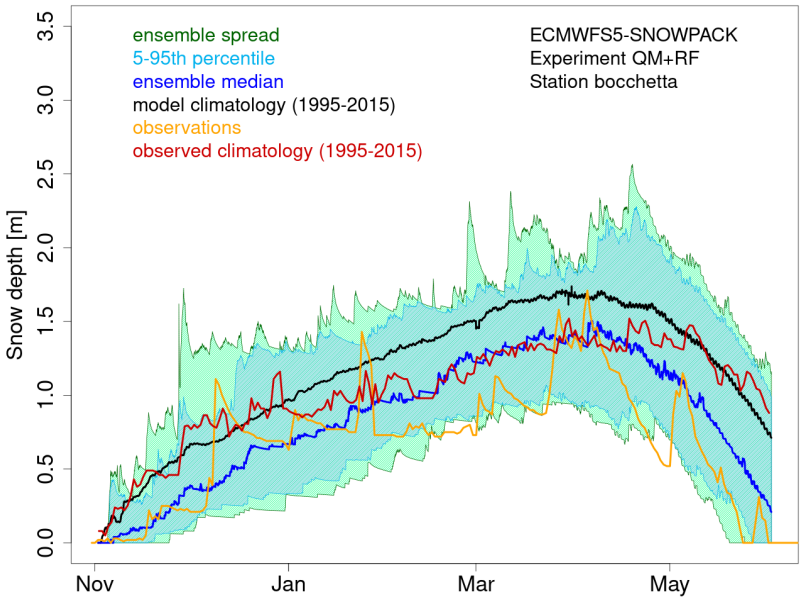

Figure 3. ECMWFS5-SNOWPACK snow depth ensemble forecasts (QM+RainFARM experiment, 250 ensemble members) initialized on November 1^{st} 2006 and issued for the 7 months ahead, for the site of Bocchetta delle Pisse (2410 m a.s.l., North Western Italian Alps). Dark green lines represent the ensemble spread, cyan lines represent the 5^{th}-95^{th} percentile range of the snow depth distribution, the blue line represents the ensemble median of the snow forecasts over the considered season, the black line represents the ensemble median of the forecasts over the reference period 1995-2015, the orange line represents in-situ observations and the red line represents the median of the observations over the reference period 1995-2015.

median of the forecasts for the season 2006/2007 are compared to the ensemble median of all forecasts for all seasons of the period 1995-2015 in order to highlight the characteristics of the considered season with respect to model climatology and determine if snow depth is expected to be below or above median. The plot also reports the snow depth observations for that season and the observed climatology to visually inspect the accuracy of the forecast (please note that differences between the observed and the modelled climatology are due to uncertainties in the bias-adjusted meteorological forcing and in the snow model structure).

We present the output of the modelling chain also in the form of tercile-based forecasts (Figure 4). For each month of the season, the tercile-based forecast plot shows the probability density function (PDF) of the 250 monthly mean snow depth forecasts, together with the probabilities to have snow depth in each tercile, and the indication of the most likely tercile. The plot also reports the probability for snow depth to be lower than the 10^{th} percentile and higher than the 90^{th} percentile. Tercile and percentile thresholds are calculated on the 21*250 monthly mean snow depth forecast values over the period 1995-2015. In the example reported in Figure 4 snow depth forecasts indicate the lower tercile (below normal) as most likely in each month of the snow season. In order to visually evaluate the quality of the forecast, the observed snow depth is also reported: if the

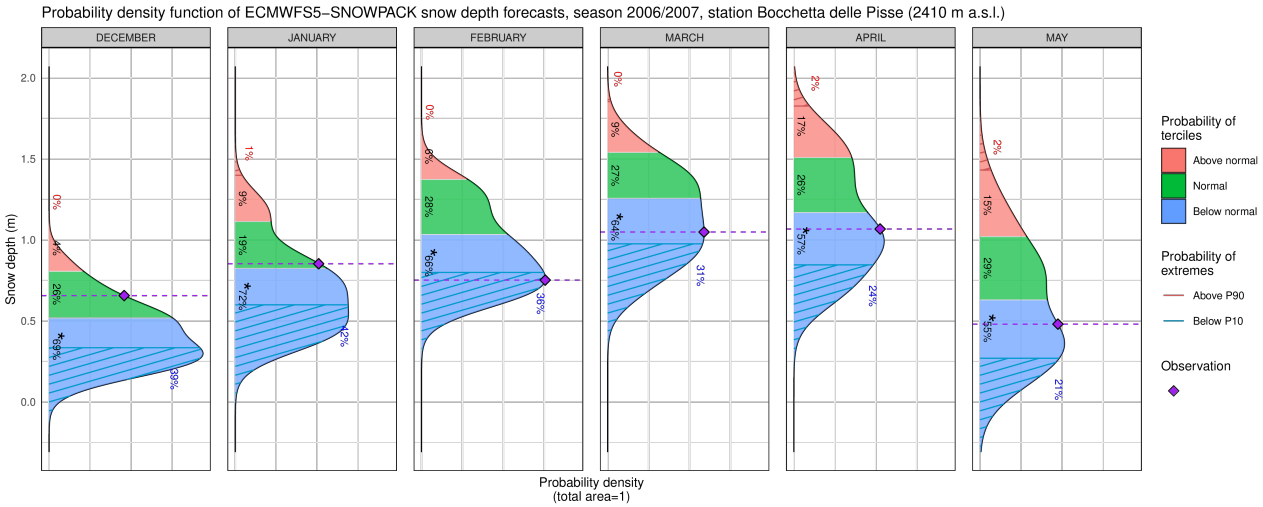

Figure 4. Probability Density Functions (PDFs) of the ECMWFS5-SNOWPACK monthly mean snow depth ensemble forecasts for the season 2006-2007 and for the station of Bocchetta delle Pisse, 2410 m a.s.l. in the Italian Alps. Areas in blue, green and coral colors represent the % probability to have monthly average snow depth below, near and above the normal conditions for the period, respectively, and the asterisk indicates the most likely tercile. Areas with blue and red parallel lines represent the probability to have monthly snow depth below the 10^{th} percentile and above the 90^{th} percentile, respectively. Observations are reported as purple diamonds.

observed snow depth falls within the most likely tercile, the forecast is successful. In this season the forecast is successful in February, March, April and May, so in late winter and spring.

3.2 Effects of the precipitation bias-adjustment and downscaling

The snow depth forecast presented in Figs. 3 and 4 is obtained after applying quite sophisticated bias correction and downscaling techniques to precipitation data. In this section we assess the added value, if any, of applying those bias-adjustment and/or downscaling methods compared to the use of raw precipitation data. We present the results of the 4 experiments (RAW, QM, RainFARM, QM+RainFARM) listed in Table 3, in which we apply or not the correction methods to precipitation forecasts. We use an indirect approach, i.e. we assess the added value of total precipitation corrections by measuring the agreement between the snow depth climatology obtained from the 4 experiments and the observed climatology in terms of root mean square error (RMSE). For each of the two forecast systems, ECMWS5 and MFS6, and each experiment, Figure 5 shows the simulated snow depth climatology (multi-annual and multi-member average) compared to the observed climatology at the station of Bocchetta delle Pisse for the period 1995-2015. Figure 5 also shows the two ERA5 snow depth climatologies obtained using raw (ERA5) and bias-corrected (ERA5_{MBC}) temperature forcing, respectively (Sect. 2.2). The corresponding RMSEs are reported in Table 4.

When SNOWPACK is driven by ERA5 forcing (raw temperature), the model RMSE on snow depth is in the range 0.30-0.35 m for Bocchetta delle Pisse and Lago Agnel stations, while it is higher (RMSE=0.5 m) for Rifugio Gastaldi: in this last station,

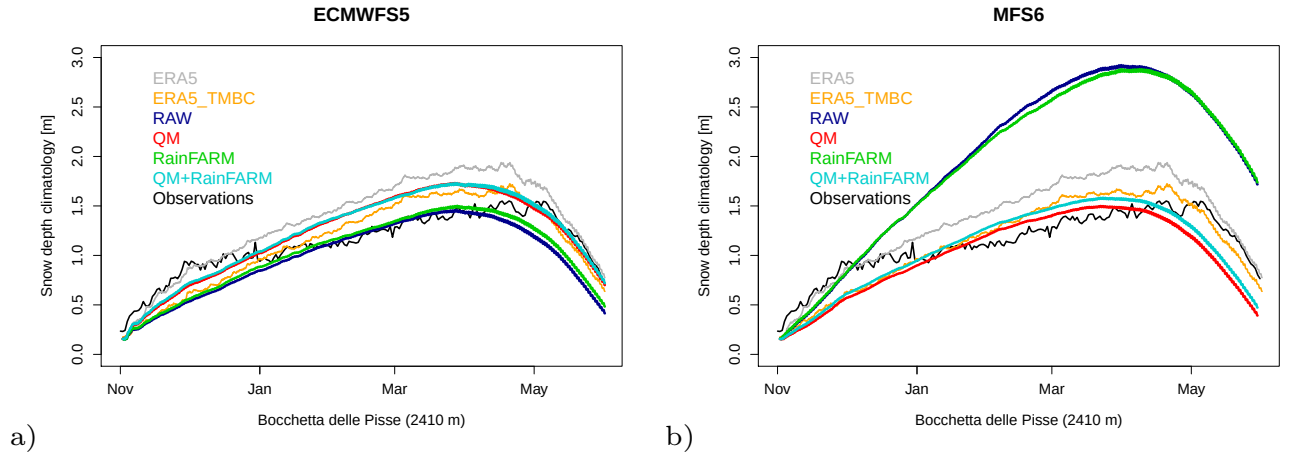

Figure 5. Daily snow depth climatology for the period 1995-2005 as simulated by the SNOWPACK model forced by ERA5 when using (gray) raw and (orange) bias-corrected air temperature, and by (a) ECMWFS5 and (b) MFS6 seasonal forecasts data with different precipitation input (RAW, QM, RainFARM and QM+RainFARM) as specified in Table 3, for the site of Bocchetta delle Pisse. Observations are reported in black for comparison.

Table 4. RMSE between simulated and observed daily snow depth climatologies at the station of Bocchetta delle Pisse for the experiments listed in Table 3. Model simulations are obtained by forcing SNOWPACK with ERA5, ECMWFS5 and MFS6 meteorological variables. ECMWFS5 and MFS6-driven experiments (RAW, QM, RF and QM+RF) differ in the treatment of total precipitation (see Table 3).

	RMSE [m]									
	ERA5	ERA5$_{TMBC}$	ECMWFS5				MFS6			
			RAW	QM	RF	QM+RF	RAW	QM	RF	QM+RF
Bocchetta delle Pisse	0.31	0.14	0.19	0.21	0.16	0.21	1.04	0.21	1.01	0.20
Rifugio Gastaldi	0.50	0.27	0.27	0.35	0.30	0.45	1.38	0.39	2.13	0.57
Lago Agnel	0.32	0.15	0.18	0.22	0.37	0.60	1.19	0.24	2.63	0.78

snowfalls are typically followed by rapid snow ablation (not shown), so the large RMSE can be related to ERA5 issues in capturing the meteorological conditions responsible for the fast melting. When bias-corrected (ERA5$_{TMBC}$) instead of raw temperature input is used, the SNOWPACK RMSE is remarkably reduced at all the three stations: the reduction is by more than 50% at Bocchetta delle Pisse and Lago Agnel, with RMSE of 0.14 and 0.15 m respectively, and by almost 50% at Rifugio Gastaldi with RMSE=0.27 m. A simple bias-correction of ERA5 temperature input is sufficient to remarkably improve the agreement between the simulated and observed snow depth climatology. ERA5-driven simulations are the reference against which to compare seasonal-forecast-driven simulations. Compared to the ERA5$_{TMBC}$ run, the RAW experiment shows similare RMSE when using the ECMWFS5 forcing and remarkably higher RMSE when using the MFS6 forcing. This suggests that

after the bias-adjustment of both ERA5 and seasonal forecast temperature i) the ECMWFS5 forcing has comparable accuracy as the ERA5 forcing; ii) the MFS6 forcing has residual systematic errors that affect the reliability of the simulations.

The application of the quantile mapping to heavily biased precipitation forecasts (MFS6) allows for a clear improvement of the model RMSE which is reduced up to almost 5 times compared to the RAW experiment. On the other hand, the application of the quantile mapping to already accurate forcing (ECMWFS5) can have different effects depending on the accuracy of the reference dataset. Here the application of the quantile mapping using ERA5 as a reference has no remarkable effects (Bocchetta delle Pisse and Lago Agnel) or it slightly increases (Rifugio Gastaldi) the RMSE (see Table 4, ECMWFS5 model, QM experiment) but it might also have detrimental effects when the reference dataset is inaccurate.

The application of the RainFARM downscaling (RF experiment) produced small effects at Bocchetta delle Pisse station (orographic weight equal to 1.05), and gradually more relevant effects at Rifugio Gastaldi and Lago Agnel (weights equal to 1.21 and 1.43, respectively, see Terzago et al. (2018) for details). In these last two cases the orographic downscaling amplifies precipitation amounts and leads to an overestimation of the snow depth output, with snow depth errors doubling for about 50% increase in the precipitation input (Lago Agnel).

These results suggest that the choice of the forecast system strongly impacts the agreement between the simulated and the observed climatology. The application of the quantile mapping is recommended in case of large biases in the precipitation input, in order to reproduce a snow depth climatology as realistic as possible. However, the application of the quantile mapping is recommended only if a trusted, reliable reference dataset is available. In fact, if the reference dataset is less accurate than the dataset that we want to correct, the application of the bias adjustment may lead to larger errors. The RainFARM downscaling is blind to model biases so, in presence of heavily biased forcing, it should be applied only after bias correction. Since the downscaling might have either positive or negative effects depending on the orographic weights, the added value of the downscaling should be checked against observations before using the fine scale precipitation data.

3.3 Evaluation of the snow depth forecasts

In order to assess the skill of the forecasting method presented in this study we evaluate the snow depth forecasts over the period 1995-2015 (hindcasts) in comparison to snow depth observations, using the set of metrics introduced in Sect. 2.7. We recall that all metrics are calculated on detrended time series.

3.3.1 Time correlation

Figure 6 shows the correlation between ensemble mean monthly and seasonal hindcasts and observations for the two seasonal forecast systems, ECMWFS5 and MFS6, and for the four experiments listed in Table 3 for the station of Bocchetta delle Pisse. Confidence intervals represented in Fig. 6 as error bars or as a gray rectangle correspond to the 5-95th percentile range of 1000 bootstrap samples derived as described in Sect. 2.7. The correlation values for all three stations, together with their significance at 95% confidence level, are reported in Table 5.

A common behavior is found among all stations: the correlation is highest in November, i.e. at lead time 1-month when the meteorological input is generally well correlated with observations, then the correlation decreases reaching a minimum

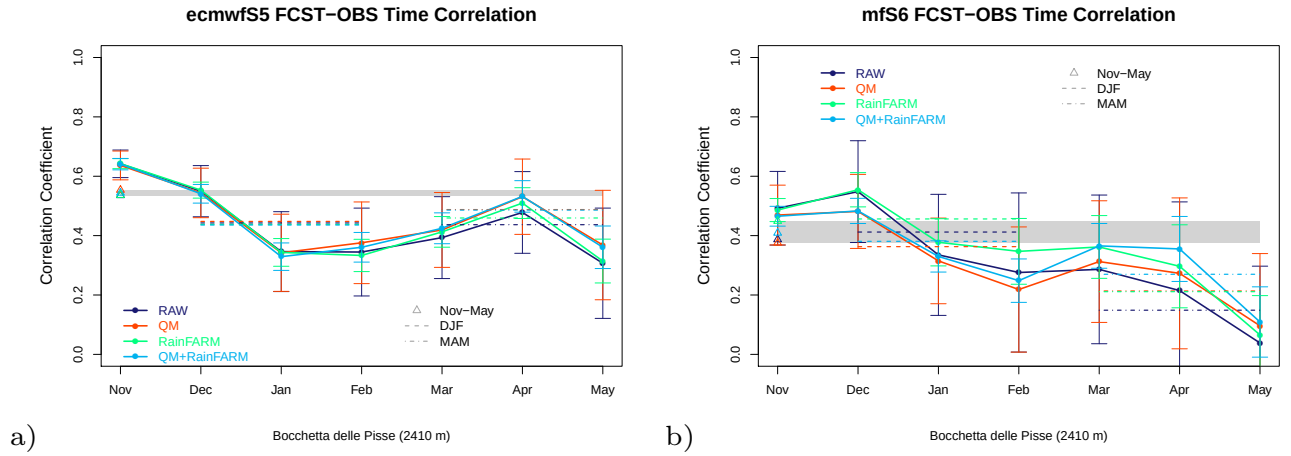

Figure 6. Pearson's correlation coefficient between forecasts of ensemble-mean monthly-mean snow depth obtained with (a) ECMWFS5 and (b) MFS6 forcing and observations at the site of Bocchetta delle Pisse. Forecasts are initialized on November 1^{st} and run with a lead time of 7 months. Colored dots represent the correlation for each month and each experiment; horizontal dashed (dash-dotted) lines represent DJF (MAM) values; the gray rectangle and the 4 colored triangles represent seasonally-averaged (Nov-May) values. Error bars represent the 5-95 th percentile range of the distribution of 1000 bootstrap samples as described in Sect. 2.7.

in winter months (January or February depending on the station and forcing). After February the correlation increases to a secondary maximum in April, then it finally drops in May. Correlation values are very similar among different experiments, especially for the ECMWFS5 model. The largest differences among experiments are found for the MFS6 model in spring (March and April), when QM and QM+RainFARM experiments provide higher time correlations than the RAW experiment, although they lie within the uncertainty range of the RAW experiment and none of these correlations is statistically significant.

Focusing on significant correlations at 95% confidence level (Table 5), we observe differences between seasonal forecast systems: using ECWMFS5 forcing, correlations are significant for all stations, all experiments and most lead times: the correlation is significant at lead time 1- and 2-month (November and December, respectively) and, interestingly, also at lead time 5- and 6-months (March and April), at the seasonal (November-May), winter (DJF) and spring (MAM) scale. Correlation is generally not statistically significant in May, and for some stations (Bocchetta delle Pisse and Lago Agnel) and experiments also in January and February. Compared to ECMWFS5, MFS6 correlation is considerably lower and generally not statistically significant after December, probably owing to a lower skill and larger biases in the meteorological forcing.

In challenging conditions such as poor meteorological forcing (MFS6) the application of bias-adjustment, downscaling or the combination of both, generally improves correlations with respect to the RAW experiment, however this improvement does not lead to statistically significant correlations.

Table 5. Time-correlation of the detrended mean monthly snow depth forecasts with respect to observations at the three stations for ECMWFS5 and MFS6 systems. Correlations significant at 95% confidence level are identified in bold and by an asterisk (*).

	Pearson time correlation														
	BOCCHETTA DELLE PISSE					**RIFUGIO GASTALDI**					**LAGO AGNEL**				
	ECMWFS5														
	ERA5	**RAW**	**QM**	**RF**	**QM+RF**	**ERA5**	**RAW**	**QM**	**RF**	**QM+RF**	**ERA5**	**RAW**	**QM**	**RF**	**QM+RF**
Nov	**0.93***	**0.64***	**0.64***	**0.64***	**0.64***	**0.44***	**0.74***	**0.76***	**0.73***	**0.75***	**0.91***	**0.62***	**0.66***	**0.65***	**0.68***
Dec	**0.92***	**0.55***	**0.54***	**0.55***	**0.54***	**0.52***	**0.59***	**0.59***	**0.59***	**0.60***	**0.89***	**0.51***	**0.53***	**0.53***	**0.53***
Jan	**0.88***	0.35	0.34	0.34	0.33	**0.71***	**0.46***	**0.45***	**0.43***	**0.45***	**0.84***	0.37	0.38	**0.39***	**0.39***
Feb	**0.84***	0.34	**0.38***	0.33	0.36	**0.71***	**0.37***	**0.38***	0.36	**0.39***	**0.89***	0.35	0.37	**0.41***	**0.43***
Mar	**0.85***	**0.39***	**0.42***	**0.41***	**0.42***	**0.67***	**0.41***	**0.39***	**0.39***	**0.40***	**0.89***	**0.43***	**0.41***	**0.50***	**0.48***
Apr	**0.79***	**0.48***	**0.53***	**0.51***	**0.53***	**0.60***	**0.40***	**0.39***	**0.41***	**0.38***	**0.92***	**0.43***	**0.44***	**0.49***	**0.46***
May	**0.80***	0.31	0.37	0.31	0.36	**0.66***	0.30	0.30	0.30	0.30	**0.93***	0.26	0.29	0.33	0.32
NM	**0.89***	**0.54***	**0.55***	**0.54***	**0.54***	**0.66***	**0.52***	**0.53***	**0.52***	**0.53***	**0.92***	**0.48***	**0.51***	**0.54***	**0.54***
DJF	**0.90***	**0.44***	**0.45***	**0.44***	**0.44***	**0.65***	**0.50***	**0.50***	**0.48***	**0.51***	**0.88***	**0.42***	**0.45***	**0.46***	**0.47***
MAM	**0.82***	**0.44***	**0.49***	**0.46***	**0.49***	**0.70***	**0.39***	**0.39***	**0.39***	**0.39***	**0.93***	**0.41***	**0.41***	**0.46***	**0.44***
	MFS6														
	ERA5	**RAW**	**QM**	**RF**	**QM+RF**	**ERA5**	**RAW**	**QM**	**RF**	**QM+RF**	**ERA5**	**RAW**	**QM**	**RF**	**QM+RF**
Nov	**0.93***	**0.49***	**0.47***	**0.49***	**0.47***	**0.44***	**0.44***	**0.45***	**0.43***	**0.44***	**0.91***	0.30	0.28	0.29	0.29
Dec	**0.92***	**0.55***	**0.48***	**0.55***	**0.48***	**0.52***	**0.46***	**0.43***	**0.45***	**0.42***	**0.89***	**0.45***	**0.40***	**0.45***	**0.41***
Jan	**0.88***	0.34	0.31	**0.38***	0.33	**0.71***	0.21	0.29	0.16	0.24	**0.84***	0.23	0.27	0.20	0.25
Feb	**0.84***	0.28	0.22	0.35	0.25	**0.71***	0.11	0.16	0.09	0.12	**0.89***	0.23	0.27	0.21	0.27
Mar	**0.85***	0.29	0.31	0.36	0.37	**0.67***	0.15	0.24	0.12	0.20	**0.89***	0.29	**0.41***	0.24	0.36
Apr	**0.79***	0.21	0.27	0.30	0.36	**0.60***	0.09	0.17	0.06	0.13	**0.92***	0.19	0.30	0.18	0.28
May	**0.80***	0.04	0.10	0.07	0.11	**0.66***	0.03	0.07	0.01	0.05	**0.93***	0.09	0.14	0.10	0.16
NM	**0.89***	**0.39***	**0.38***	**0.45***	**0.41***	**0.66***	0.22	0.29	0.18	0.25	**0.92***	0.28	0.32	0.25	0.33
DJF	**0.90***	**0.41***	0.36	**0.46***	**0.38***	**0.65***	0.27	0.31	0.23	0.28	**0.88***	0.31	0.33	0.29	0.33
MAM	**0.82***	0.15	0.21	0.21	0.27	**0.70***	0.06	0.14	0.04	0.10	**0.93***	0.16	0.26	0.15	0.24

3.3.2 Brier Skill Score

The Brier Skill Score (BSS) shows the relative skill of the forecast prototype with respect to the climatological forecast in terms of mean square error of the probability forecasts for a binary event. In our case the binary event is "snow depth in a given tercile of the forecast distribution". BSS takes positive values whenever the forecast prototype is more skillful than climatology. Figure 7 shows the time evolution of BSS for the two seasonal forecast systems, ECMWFS5 and MFS6, and for the four experiments listed in Table 3 for the station of Bocchetta delle Pisse. Error bars computation is based on 1000 bootstrap samples derived as described in Sect. 2.7. The winter (DJF), spring (MAM) and seasonal (Nov-May) BSS values are reported in the plot as dashed lines, dot-dashed lines, and grey strips respectively. BSS values for all three stations are reported and compared in Figure 8, where positive (negative) BSSs are highlighted in hues of green (blue) color, and a discretized scale with

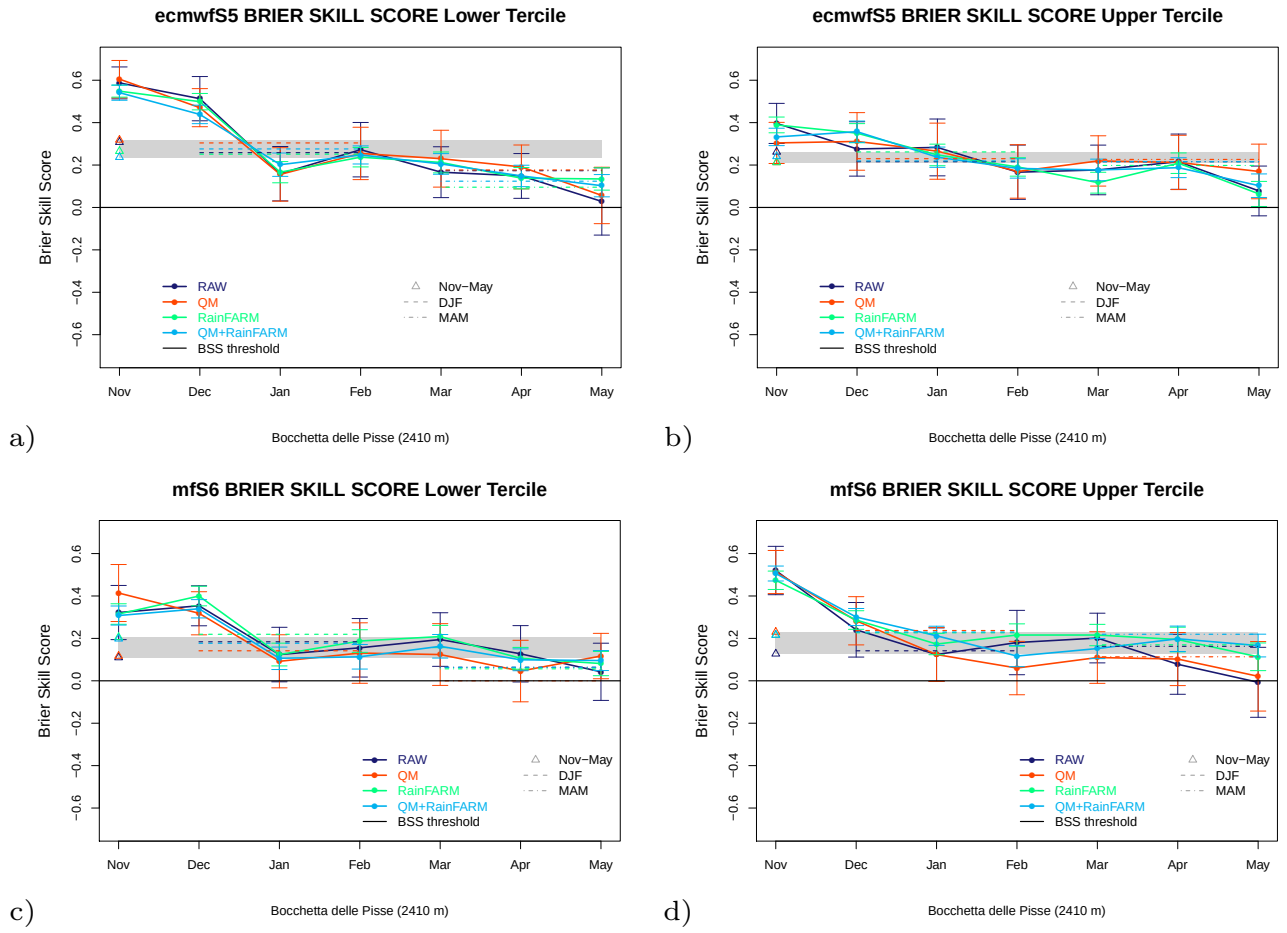

Figure 7. Brier Skill Score for seasonal forecasts of monthly- and seasonally-averaged snow depth in the (a,c) lower and (b,d) upper terciles, for (a,b) ECMWFS5 and (c,d) MFS6 forcing, starting date November 1^{st}, lead times from 1 to 7 months for the site of Bocchetta delle Pisse. Colored dots represent the BSS for each month and each experiment; horizontal dashed (dash-dotted) lines represent DJF (MAM) BSS values; the gray filled rectangle and the 4 colored triangles all refer to the seasonal (Nov-May) values, indicating the BSS spread (min-max) and the single BSS values for the 4 experiments, respectively.

thresholds of 0, ± 0.2, ± 0.4 allows to distinguish between *fair*, *good*, and *remarkable* skill, respectively (i.e. *fair* corresponds to $0 < BSS \leqslant 0.2$, *good* corresponds to $0.2 < BSS \leqslant 0.4$, *remarkable* corresponds to $BSS > 0.4$).

The BSS is generally positive for both seasonal forecast systems, both lower and upper terciles, for almost all experiments, all lead times and all stations (Figs. 7 and 8). The BSS is highest in November and/or December and then it decreases reaching its minimum, but still with positive values (in all cases but one close to zero) in May (Fig. 7), demonstrating a clear added value of the prototype forecast with respect to the climatological forecast. ECMWFS5 generally shows higher BSS than MFS6 for both lower and upper terciles, indicating better forecast skills than its counterpart. The difference is more evident in DJF when ECMWFS5 shows predominantly good or even remarkable skill, while MFS6 shows predominantly good or fair skill.

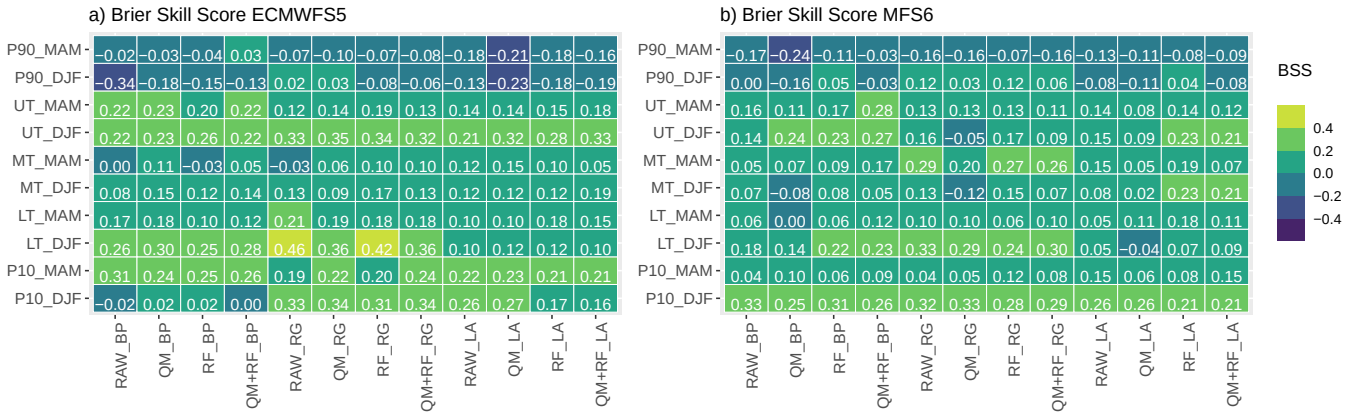

Figure 8. Brier Skill Score of the detrended seasonal (DJF, MAM) snow depth forecasts in the lower (LT), middle (MT), upper (UT) tercile, as well as in the lower (P10) and upper (P90) extreme of the distribution, with respect to the climatological forecasts, using observations at the three stations as a reference, for a) ECMWFS5 and b) MFS6 systems. Positive and negative BSSs are highlighted in shades of green and blue, respectively.

In MAM ECMWFS5 still outperforms MFS6 but the difference between the two is reduced, and both show fair skill in most of experiments. MFS6 shows larger differences between the four experiments, without a clear relation between the prototype skill and the application of the bias-adjustment and downscaling methods to precipitation data.

3.3.3 Area Under the ROC curve Skill Score

AUCSS is a measure of the "discrimination" of the seasonal forecast system: it indicates how good are individual hindcasts at discriminating mean monthly snow depth falling in the upper, middle and lower tercile in comparison to the reference climatological forecast. We recall that positive values indicate improvements, while negative values indicate poorer skills than the reference climatological forecast. Figure 9 shows the time evolution of AUCSS for the two seasonal forecast systems, ECMWFS5 and MFS6, for the four experiments listed in Table 3, for the station of Bocchetta delle Pisse and for the lower and upper terciles. Error bars are calculated based on 1000 bootstrap samples derived as described in Sect. 2.7. The winter (DJF) and spring (MAM) AUCSS values for all three stations are reported in Figure 10, where positive (negative) AUCSS are highlighted in greenish (bluish) colors.

Considering the ECMWFS5 forecasting system, a clear added value emerges in predicting the events in the terciles below normal and above normal for all stations, all experiments and all lead times at least up to April included (up to May for Lago Agnel, not shown). For all stations the AUC skill scores at the seasonal scale (DJF and MAM) indicate an improvement with respect to the climatological forecast, with remarkable forecast skill in winter and generally good skill in spring.

Considering the MFS6 forecast system, we find a clear added value at forecasting snow depth in the upper tercile (generally with good or remarkable skills in DJF and a more or less strong decrease in MAM) and in the lower tercile in DJF. The prediction skills for MAM snow depth in the lower tercile depend on the station: in detail, skills are good or remarkable for

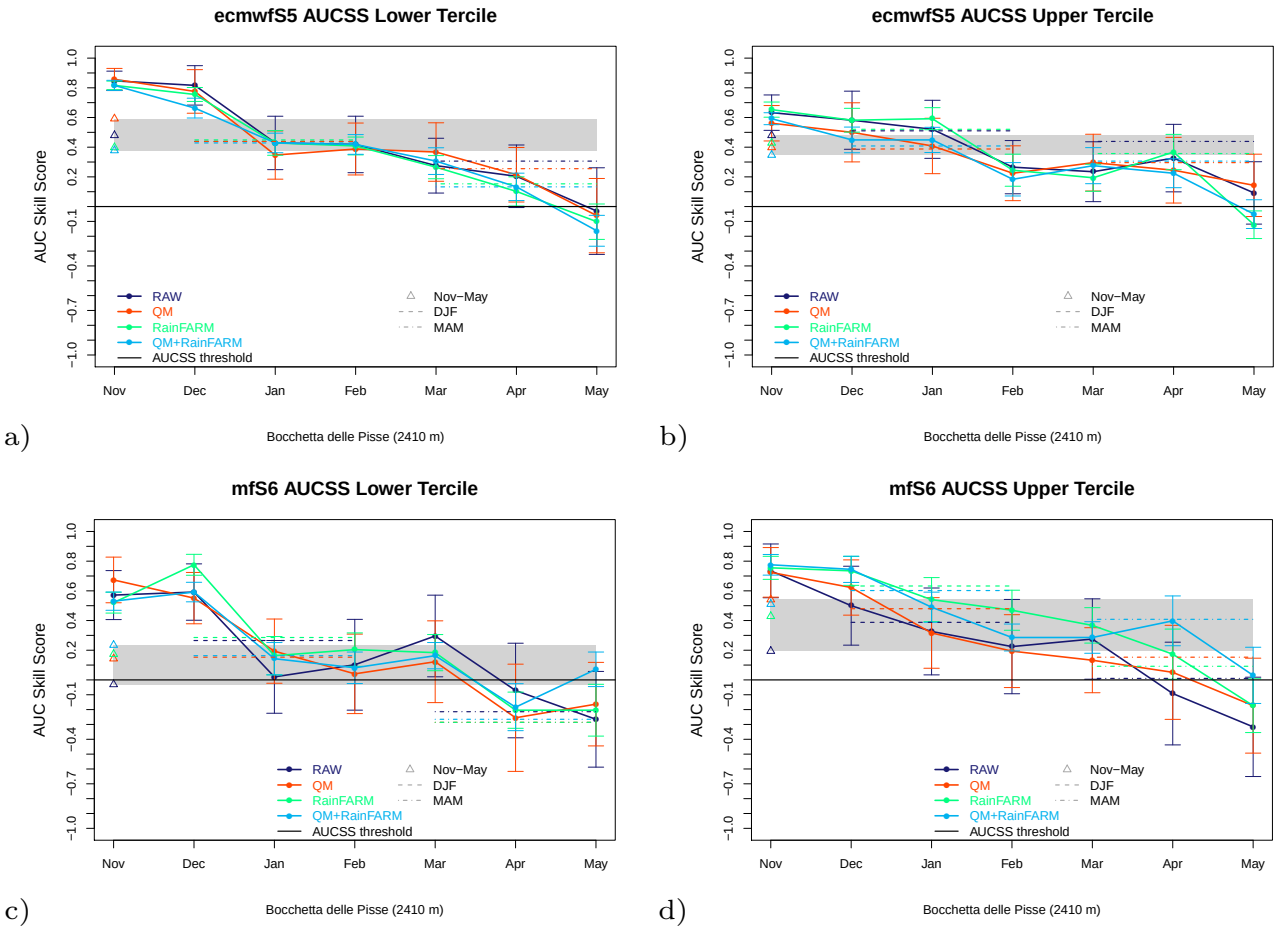

Figure 9. AUCSS for seasonal forecasts of monthly- and seasonally-averaged snow depth in the (a,c) lower and (b,d) upper terciles, for (a,b) ECMWFS5 and (c,d) MFS6 forcing, starting date November 1^{st}, lead times from 1 to 7 months for the site of Bocchetta delle Pisse. Colored dots represent the AUCSSs for each month and each experiment; horizontal dashed (dash-dotted) lines represent DJF (MAM) AUCSS values; the gray filled rectangle and the 4 colored triangles all refer to the seasonal (Nov-May) snow depth forecasts, indicating the AUCSS spread (min-max) and the AUCSS values for the 4 experiments, respectively.

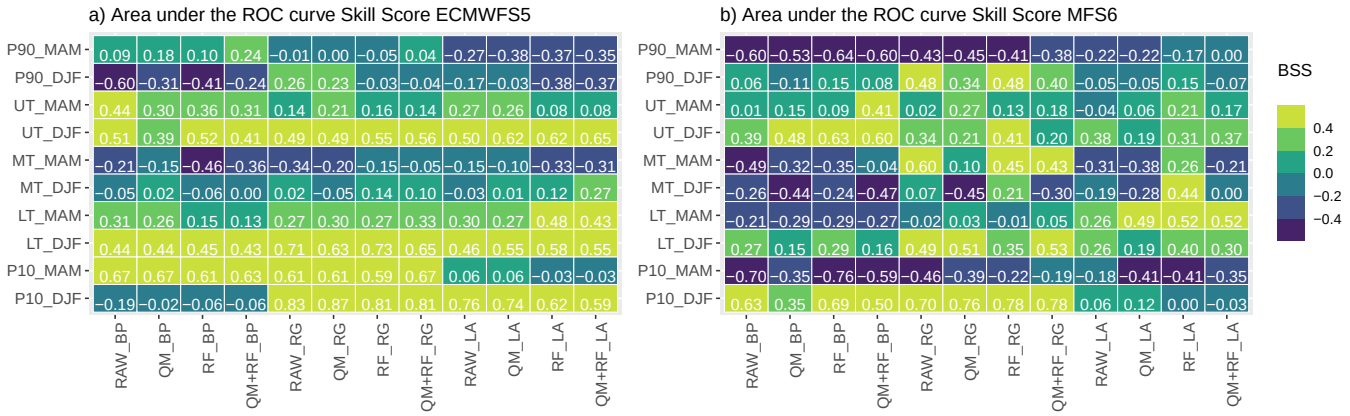

Figure 10. AUCSS of the detrended seasonal (DJF, MAM) snow depth forecasts in the lower (LT), middle (MT), upper (UT) tercile, as well as in the lower (P10) and upper (P90) extreme of the distribution, with respect to the climatological forecasts, using observations at the three stations as a reference, for a) ECMWFS5 and b) MFS6 systems. Positive and negative AUCSSs are highlighted in shades of green and blu, respectively colors, respectively.

Lago Agnel, while contrasted results with both positive and negative skills depending on the experiment are found for Rifugio Gastaldi, and negative skills are found for Bocchetta delle Pisse.

Seasons with snow depth within the norm are usually predicted with similar or lower skills than the climatological forecast, with some differences depending on the seasonal forecast system. While ECMWFS5 shows limited added value in all stations, experiments and seasons, the skill of MFS6 is more station and experiment dependent, and some skill is found for Rifugio Gastaldi and Lago Agnel stations (see Fig. 10 for more details).

It is interesting to note that limited to the upper tercile, the AUCSS generally shows a secondary maximum in March or April (particularly evident for Rifugio Gastaldi and Lago Agnel stations, not shown) indicating that the forecast system has skills at predicting spring seasons with above normal snow depth. For the lower tercile this secondary maximum is often less pronounced.

The largest differences among the four experiments are found for MFS6, however there is not a single experiment usually performing better than others.

3.3.4 Continuous ranked probability score (CRPS)

The continuous ranked probability score (CRPS) is a measure of the overall accuracy of the ensemble forecast. The Brier score and the CRPS are complementary measures, with the former providing information on the accuracy of tercile-based forecasts and the latter evaluating the overall accuracy of the forecast distribution, considering the entire permissible range of values for the considered variable. Figure 11 shows the time evolution of CRPSS for the two seasonal forecast systems, ECMWFS5 and MFS6, and for the four experiments listed in Table 3 for the station of Bocchetta delle Pisse. In addition to the plots, Figure 12

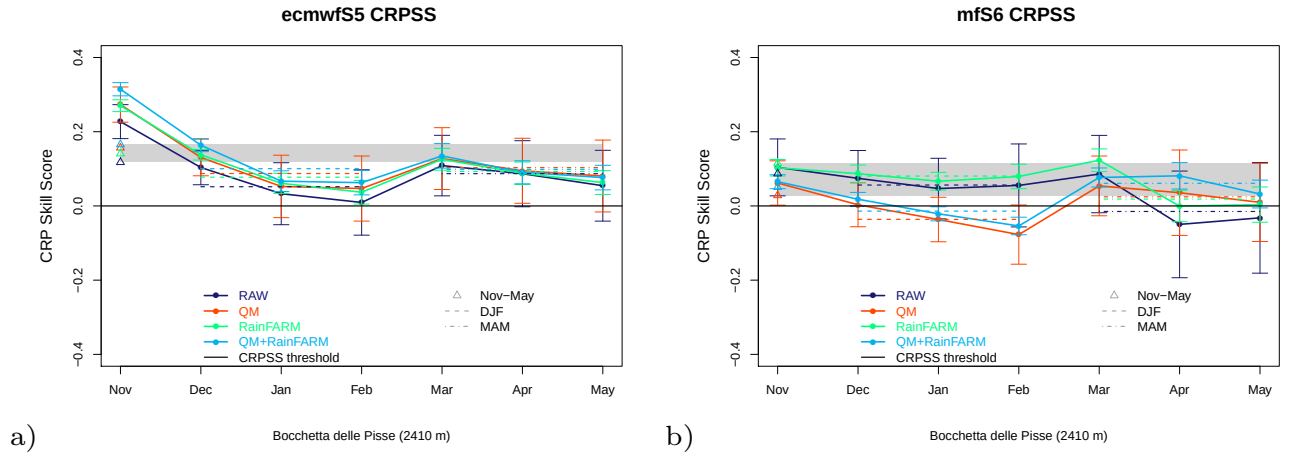

Figure 11. CRPSS for seasonal forecasts of monthly- and seasonally-averaged snow depth for (a) ECMWFS5 and (b) MFS6 forcing, starting date November 1^{st}, lead times from 1 to 7 months, for the site of Bocchetta delle Pisse. Colored dots represent the scores for each month and each experiment; horizontal dashed (dash-dotted) lines represent DJF (MAM) scores; the gray filled rectangle and the 4 colored triangles all refer to the seasonal (Nov-May) snow depth forecasts, and they indicate the score spread among the 4 different experiments and the score for each of the 4 experiments, respectively.

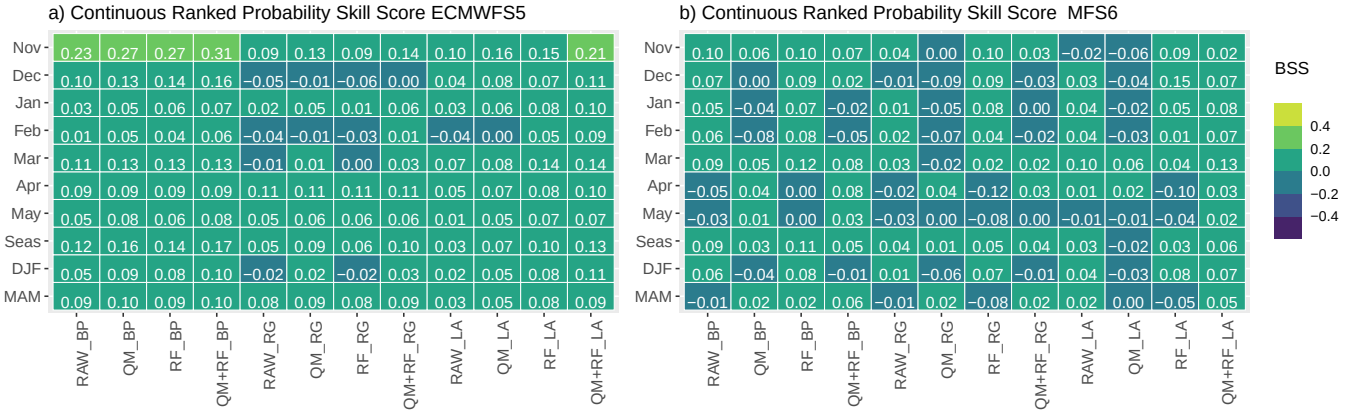

Figure 12. CRPSS of the detrended monthly and seasonal (DJF, MAM) snow depth forecasts with respect to the climatological forecasts, using observations at the three stations as a reference, for a) ECMWFS5 and b) MFS6 systems. Positive and negative CRPSSs are highlighted in shades of green and blu, respectively colors, respectively.

shows the monthly and seasonal CRPSS values for all three stations, with positive (negative) CRPSS are highlighted in shades of gree (blu).

Considering the ECMWFS5 forecasting system, the CRPSS is generally positive, although with small values, across the different experiments, lead times and most of stations. Few exceptions with CRPSS values close to zero are found, and they are mostly in winter months. When using the MFS6 forecasting system, the skill is lower than ECMWFS5: the skill is present up

to lead times 5-months (March) only for selected stations and experiments. Skills at lead time 6-7 months (April, May) and/or in the QM experiment are rare, suggesting a worsening of the performances when the total precipitation input is bias-adjusted with the quantile mapping method with respect to ERA5. The application of the quantile mapping with respect to ERA5 seems to waste some of the limited forecast skill.

Like many other skill scores analyzed, also the CRPSS decreases from November up to the end of the winter, then it increases again for a secondary maximum in March or April. This behavior is very common and it seems a robust feature across different forecast systems, experiments and test sites. Overall, the presence of positive CRPSS values, also when the score reaches its minimum, clearly indicates the added value of the prototype forecast than the climatological forecast, in terms of overall accuracy.

3.3.5 Events outside the 10-90th percentile range

The analysis of the prototype performance also covers the ability to predict events below the 10^{th} percentile (P10, lower extreme) and above the 90^{th} percentile (P90, upper extreme). Figure 13 shows the time evolution of BSS for extreme values for the two seasonal forecast systems, ECMWFS5 and MFS6, and for the four experiments listed in Table 3 for the station of Bocchetta delle Pisse. Figure 8 summarizes the BSS values for all the stations. Looking at the plots for Bocchetta delle Pisse station for the events below P10 (Figure 13a, 13c), the BSS is generally positive during the snow season, indicating a clear skill at predicting low snow months/seasons. In only one case the BSS is close to zero in all experiments (i.e. EMWFS5 forcing, DJF season, Bocchetta delle Pisse station) and the application of bias correction, downscaling or the combination of both do not improve the skill. In all other cases, the skill is robust across different forecast systems, seasons, experiments and stations. It is interesting to note that MFS6 shows good skills at forecasting months/seasons with snow below P10, with similar performances or even outperforming the ECMWFS5-driven experiments. Looking at the plots for Bocchetta delle Pisse station for the events above P90 (Figure 13b, 13d), the BSS is generally negative, indicating no skill of the forecast system at predicting months/seasons with exceptionally abundant snow depth. This property is maintained considering different driving models, seasons, experiments and stations. Some skill (positive BSS) are found for Rifugio Gastaldi station (all experiments) and especially when using the MFS6 forcing.

4 Discussion

In this paper we present an original prototype for generating multi-system ensemble seasonal forecasts of snow depth at the local scale from November up to May of the following year (7 months lead time), providing information which are relevant for economic activities such as hydropower production, water management and winter ski tourism. The prototype is based on the SNOWPACK model forced by meteorological data of the Copernicus Climate Data Store seasonal forecast systems, namely ECMWFS5 and MFS6. The skill of the prototype has been assessed using different deterministic and probabilistic metrics: i) the time correlation of the ensemble mean snow depth forecast with the observed snow depth; ii) the accuracy (BSS) and the

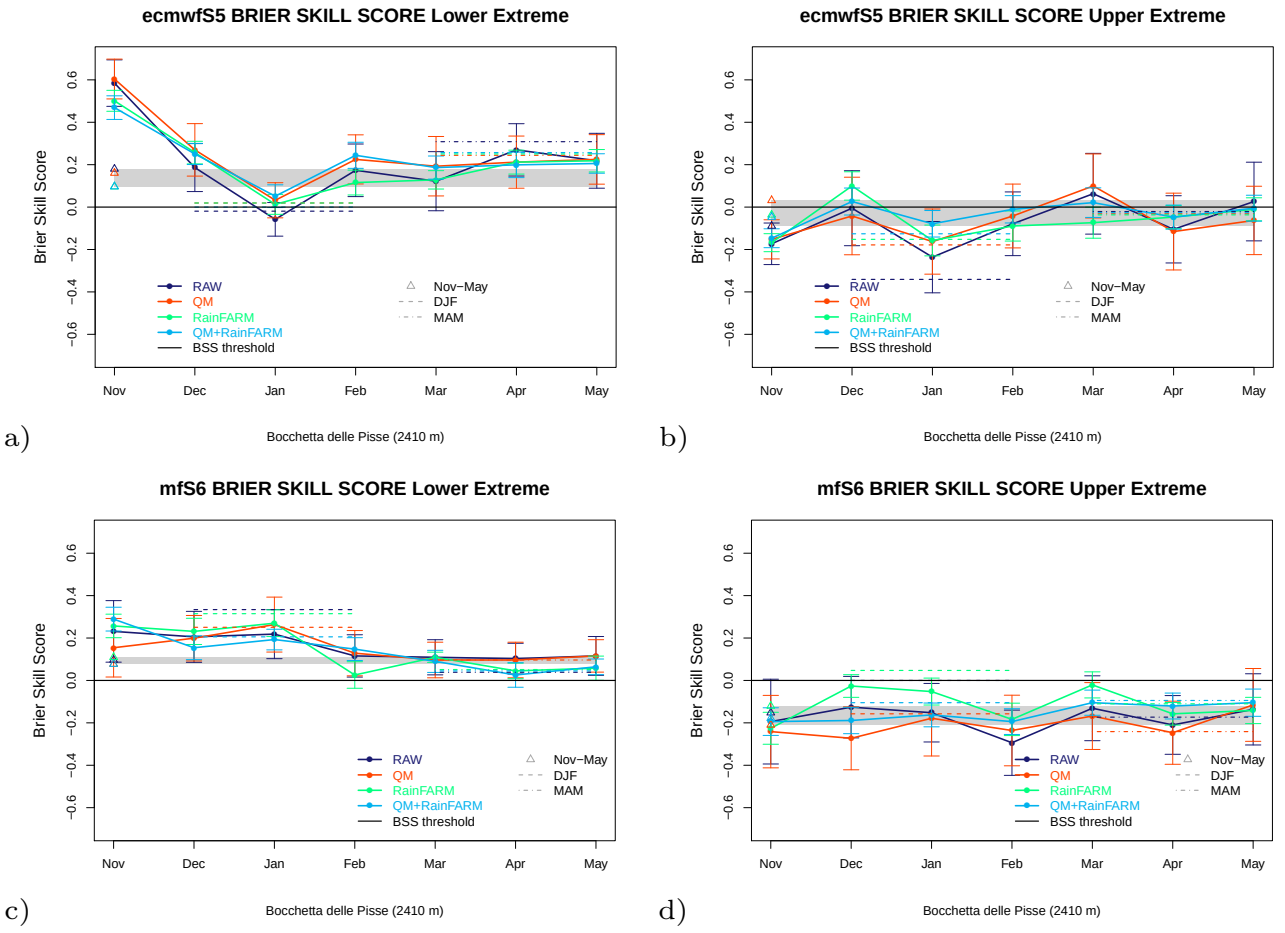

Figure 13. Brier Skill Score for seasonal forecasts of monthly- and seasonally-averaged snow depth (a,c) below the 10^{th} percentile (P10) and (b,d) above the 90^{th} percentile (P90), for (a,b) ECMWFS5 and (c,d) MFS6 forcing, starting date November 1^{st}, lead times from 1 to 7 months, for the site of Bocchetta delle Pisse. Colored dots represent the BSS for each month and each experiment, horizontal dashed (dash-dotted) lines represent DJF (MAM) BSS values; the gray filled rectangle and the 4 colored triangles all refer to the seasonal (Nov-May) snow depth forecasts, and they precisely indicate the BSS spread among different experiments and the BSS values for each of the 4 experiments, respectively.

discrimination (AUCSS) of the tercile-based forecasts; iii) the accuracy of the forecast distribution (CRPSS). All probabilistic skills have been calculated with respect to a simple forecast method based on the climatology (reference).

The prototype shows clear skill in tercile-based forecasts, i.e. higher accuracy (BSS) and higher discrimination (AUCSS) at forecasting events below and above normal compared to the climatological forecasts, independently of the driving seasonal forecast system, station, season and experiment considered. The prototype also shows skill at forecasting extreme snow seasons with snow depth below the 10^{th} percentile, while it has difficulties in predicting extremely snowy seasons (snow depth above the 90^{th} percentile).

The choice of the forecast system has an impact on the skill of the prototype, with ECMWFS5 providing more robust skill across different seasons, metrics, and experiments than MFS6. The ECMWFS5-driven prototype provides high and significant time correlation between ensemble mean snow depth forecasts and observations for different time aggregations of the forecasts, i.e. over the whole period November-May, at the seasonal scale (DJF, MAM), or even at the monthly scale in November, December, March April. These features are valid for all the three stations considered, and single stations provide even better results, with high and significant correlations also in January and February. By contrast, MFS6 shows significant correlation only at short lead times, i.e. November and/or December. The ECMWFS5-driven prototype shows skill at predicting the snow depth forecast distribution (CRPSS) at the November-May and MAM scale (all stations) and at DJF scale (for two out of three stations). On the contrary, MFS6 shows CRPSS values close to zero or slightly positive with a scattered pattern depending on the station, season and experiment. In conclusion, compared to ECMWFS5, the MFS6 forcing prototype provides less widespread skills, and the performances are more score-, season-, experiment- and station-dependent.

A common feature of both driving systems is their better skill at predicting above- or below-normal snow depth compared to near-normal snow depth. This issue has been found in several previous works (e.g. Calì Quaglia et al., 2021; Athanasiadis et al., 2017) and it has been explained with the difficulty at predicting small rather than large amplitude anomalies.

A second common feature of the two seasonal forecast systems is the time evolution of the monthly correlation: as expected it is maximum at the beginning of the season and then it decreases, however, surprisingly, it increases again to a secondary maximum in April (or March). This feature can be probably related to the fact that the spring snowpack is determined by the climatic conditions over the previous months, and even modest skill in the prediction of the main meteorological drivers (temperature and precipitation) at short lead time are reflected in skill at predicting snowpack at longer lead times. So even if temperature and precipitation forecasts do not match the corresponding observations at the monthly scale, they can match at a longer (seasonal) scale and allow for surprisingly good predictability of the snow accumulation. Moreover, enhanced climate predictability in winter due to teleconnections such as the North Atlantic Oscillation (Lledó et al., 2020) may increase the skill in forecasting snowpack in the following spring. Increasing agreement from mid-winter to spring has been found not only for the time correlation but also for other skill scores, although in this last case the signal is not consistent throughout all forecast systems, terciles and experiments.

A third common feature of the two seasonal forecast systems is their skill at forecasting extremely low snow seasons, with snow depth below the 10^{th} percentile. This result is in line with previous studies on tercile- or quintile-based streamflow prediction (Santos et al., 2021; Wanders et al., 2019) where some reliability is achieved in the lower tercile, for high forecast

probabilities. In contrast, for the upper tercile and even clearer for the middle tercile, no reliability is found. Our findings shows that it is relatively easier to predict low-snow than high-snow seasons: this feature is of key importance since the most relevant feature requested by end-users to be available from the prototype is the capability of anticipating the occurrence of low snow seasons.

The accuracy of seasonal snow forecasts is subject to multiple sources of uncertainty, which are present in the various components of the production chain, that are: forecasts of the meteorological forcing, bias adjustment methods, downscaling techniques, snow model employed, model setup and initialization. Consequently, each component has to be evaluated to assess its relative contribution to the overall forecasting accuracy.

4.1 The impact of the choice of the seasonal forecast system

At the time when our snow depth forecast prototype was developed only two seasonal forecast systems provided all the variables necessary to drive the snow model, namely ECMWFS5 and MFS6, so we considered these two. Of course, additional seasonal forecasts systems should be analyzed as soon as data become available, investigating also the skill of the multi-system ensemble compared to the models taken individually. From our results based on ECMWFS5 and MFS6, the choice of the seasonal forecast system strongly impacts the skill of the prototype in terms of time correlation between forecasted and observed snow depth, which is higher, significant and more widespread during the snow season when using ECMWFS5 forcing with respect to MFS6 forcing. The choice of the seasonal forecast system also impacts the ability of the prototype to provide forecast distributions close to the observed ones (CRPSS). However, the choice of the forecast system does not substantially affect the ability of the system at providing skillful tercile-based forecasts (BSS and AUCSS). This finding suggests that even heavily biased seasonal forecast systems such as MFS6 over the study area can provide skillful tercile-based snow depth forecasts. In a recent study a similar behavior has been found for ECMWFS5 and MFS6 DJF temperature and precipitation forecasts over the Mediterranean region (Calì Quaglia et al., 2021).

4.2 The impact of precipitation bias correction

Accurate temperature and precipitation data are essential for simulating snow processes since the former controls the phase of precipitation and snow melt, and the latter controls snow accumulation. To adjust temperature biases we employed the most accurate data available, i.e. measurements at the meteorological station, to correct the annual cycle of the seasonal forecast systems, to make it similar to the observed one. The adjustment of precipitation biases deserves more sophisticated techniques. Precipitation measurements in mountain areas are affected by large errors owing to wind drift and inadequacy of unheated and insufficiently-heated pluviometers, both leading to a large underestimation (Kochendorfer et al., 2017a, b). Clearly the lack of reliable ground measurements hampers the possibility to accurately bias-adjust seasonal forecast precipitation data. In this study we adjusted precipitation forecasts with the quantile mapping method using ERA5 reanalysis as a reference data, assuming ERA5 to be an adequate approximation of the ground truth. An alternative option would have been to estimate total precipitation from snow depth station measurements by using the parameterization included in the SNOWPACK model (Mair et al., 2013). We tested this procedure and derive total precipitation at the three stations by running the SNOWPACK model

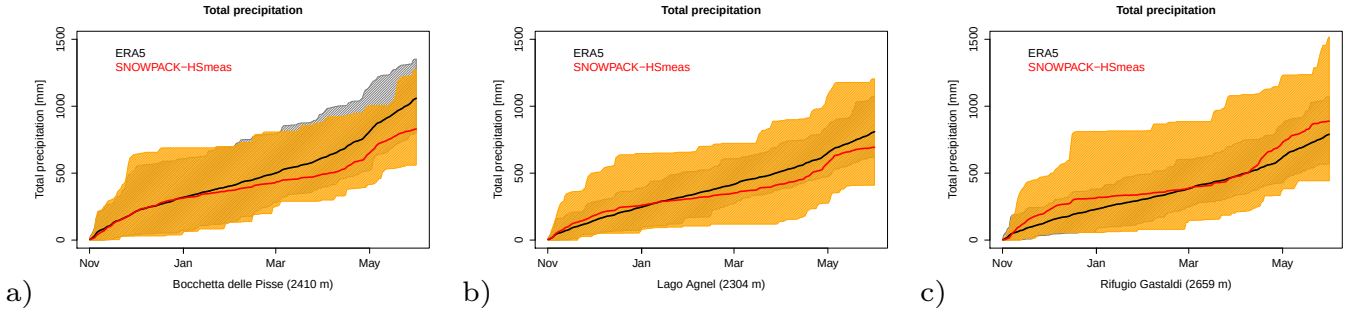

Figure 14. November-May accumulated total precipitation as estimated by (black) the ERA5 reanalysis and (red) the SNOWPACK model driven by the ERA5 forcing (all variables except for total precipitation) and the measured snow depth, for all the three stations.

driven by the ERA5 forcing (all variables used in the ERA5 experiment except for total precipitation) and the measured snow depth. We then compared the total precipitation simulated in this way to ERA5 total precipitation in terms of November-May accumulated precipitation, and the results are shown in Figure 14. At the end of May the % difference between the SNOWPACK simulated values and the corresponding ERA5 values is -22, -14 and +12% for Bocchetta delle Pisse, Lago Agnel and Rifugio Gastaldi, respectively, so it is relatively small in all the three stations. The study of how the difference between the two precipitation estimates affects the bias correction of seasonal forecasts is beyond the scope of this study and is left for further investigation. However, from the analysis carried out in this paper it is relatively easy to measure the added value of the precipitation bias correction on the simulated snow depth (Fig. 5). The precipitation adjustment is of little usefulness in case of small bias in the forecast system (ECMWFS5) when the application of the bias correction can lead to similar or slightly higher RMSEs compared to the use of RAW precipitation data. On the contrary, the application of bias adjustment to original precipitation data is useful, or even necessary, in case of strong biases in the forecast system (MFS6): in this case it allows to reconstruct the observed snow depth climatiology. In any case, however, the difference in skill scores between RAW and QM experiment is generally very small. In fact, the scores of the QM experiment lie within the range of uncertainty of the score of the RAW experiment, so the bias adjustment does not substantially influence the skills of the prototype. These results are in agreement with a former study which found that the application of the quantile mapping to seasonal forecast products eliminates forecast biases in the reforecasts, without adding much to correlation skill (Becker, B. D., 2019).

4.3 The impact of the spatial downscaling of precipitation

The application of the RainFARM downscaling to precipitation seasonal forecasts has different effects on the model RMSE depending on the station but not on the forecast system considered. In fact, the successful application of the RainFARM method (i.e. lower RMSE in the RF experiment compared to the RAW experiment) mainly depends on the accuracy of the reference climatology used to derive the weights. If the reference climatology over- or under-estimates the impact of topography on local precipitation amounts this feature will be reflected also in the downscaled data, irrespectively of the seasonal forecast system employed. So, a locally inaccurate reference climatology introduces an additional source of error (see for example the case of

Lago Agnel station, RF vs. RAW experiments). Since the results are station-dependent, we recommend checking the effects of the precipitation downscaling by verifying the improvement of the agreement between the simulated and the observed snow depth climatologies. If results are not good one should consider either using another reference dataset with higher accuracy or directly employing the original (RAW) precipitation at the coarse scale as input for the modeling chain. In support of this last option are the results of the deterministic metrics, which do not show a significant increase in skill scores when using downscaled data compared to original coarse scale data.

4.4 Spatial downscaling of other input variables

Apart from air temperature and precipitation the other variables necessary to drive the SNOWPACK model are critical to be adjusted and/or downscaled mainly due to the lack of i) surface observations to be used as a reference for bias-adjustment and ii) robust downscaling methods with proven effectiveness. Different methods have been developed to downscale wind fields, based on cluster analysis (Mengelkamp et al., 1997; Salameh et al., 2009) or using a dynamical-statistical approach (Pryor and Barthelmie, 2014), but all of these are affected by large uncertainties (Pryor and Hahmann, 2019). Martinez-García et al. (2021) shows a comparison of different statistical methods, demonstrating the non-existence of an optimal approach for all regions and applications. Humidity variables are rarely considered by downscaling studies. The most common approach consists in usage of a stepwise multiple linear regression (Anandhi, 2011). The downscaling performance depends on predictors selection, however upper air humidity variables are assessed as the most efficient ones. Spatial downscaling for incoming radiation is more complex than other variables. For example, Gupta and Tarboton (2016) downscaled MERRA reanalysis data of incoming shortwave radiation by interpolating them from coarse grid to DEM elevation one, while the incoming shortwave radiation is estimated from air temperature, cloud cover and atmospheric emissivity. In that case, the downscaling did not reduce the uncertainty of raw data. Since bias-adjustment and downscaling techniques for variables other than temperature and precipitation are affected by large uncertainties, we preferred to i) verify the overall agreement between seasonal forecasts and corresponding station measurements (when available) or ERA5 data over the period of study; provided acceptable agreement between the forecast and the reference dataset, ii) downscale seasonal forecasts using a simple bilinear interpolation to the coordinates of the station, which procedure is acceptable in absence of more sophisticated methods. Further work should clarify the effect of using more sophisticated bias-correction and donwscaling methods in the modelling chain and in particular their impact on the quality of the snow depth forecasts.

4.5 Impact of the choice of the snow model

A variety of snow models with different degrees of complexity have been developed for different purposes and applications, from very simple empirical models (e.g. degree-day models) to sophisticated, multi-layer physical snow models. An advantage of simple snow models is the limited input data requirement, which avoids uncertainties associated with other forcings, and the low computational load of the simulations. However, a limitation of simple degree-day models is that they need to be calibrated over each study site, so sufficiently long time series of forcing and validation data are necessary to calibrate and validate the model over independent subperiods. Such long term datasets are often unavailable, especially in remote areas. On the other

hand, sophisticated snow models have higher input data requirements and higher computational load compared to simpler snow models but they have the advantage that they can be directly used without calibration and their snow estimates usually have higher accuracy (Terzago et al., 2020). The choice of the appropriate model complexity depends on the objective of the work. Förster et al. (2018) aims at forecasting February SWE anomalies spatially-averaged at the catchment scale, so they employed a simple hydrological snow model driven by air temperature and precipitation anomalies only, at coarse (monthly) time resolution. Our objective is to look with finer spatial detail, moving from the catchment scale to the local scale, and forecast monthly snow depth at specific sites of interest for economic activities. In this paper we adopted a sophisticated, physical, multi-layer snow model (SNOWPACK) which provides accurate daily snow depth estimates (RMSE= 0.10 m; BIAS=0.00, Pearson-Correlation=0.79 in NW Italian Alps) across a number of different conditions and seasons (Terzago et al., 2020). The high level of accuracy of this model allows us to make the hypothesis that the model error is neglectable compared, for example, to the error associated with the forcing. This hypothesis simplifies the interpretation of the results and allows to better distinguish the contribution of the different elements of the modelling chain to the total error. The main drawback of using SNOWPACK is the number of input variables needed to run the simulations, that also limited the number of seasonal forecast systems that can be considered in this analysis.

4.6 Uncertainty in the validation data

The snow depth data used to evaluate snow forecasts are quality-controlled in-situ measurements, whose typical errors are on the order of few centimeters. This approach allows to reduce the uncertainty associated the reference data compared to more common cases in which reference data are simulated by hydrological models and model errors affect the quality of the reference data (i.e. Förster et al., 2018).

4.7 Computational costs

The modelling framework presented in this study is quite complex and includes the following steps: i) download of ensemble seasonal forecast forcing; ii) bias adjustment of temperature and precipitation; iii) spatial downscaling (all variables); iv) temporal downscaling (all variables); v) SNOWPACK simulations; vi) post-processing of the SNOWACK forecasts; vii) generation of the plots; viii) update of the website. The most time-consuming steps are the bias-adjustment and the downscaling of the precipitation input. The bias-adjustment with the quantile mapping method can substantially improve the agreement between the modelled and the observed climatology, however it is found to have a small impact on the forecast skills, especially regarding tercile-based forecasts. The limited added value of precipitation bias adjustment and downscaling to the forecast skill seems to suggest that, in these sites and in these conditions, original RAW precipitation input can be employed obtaining similar results as in the more complex frameworks.

5 Conclusions

The paper presents first-of-their-kind multi-system ensemble seasonal forecasts of the snow depth evolution from November up to May of the following year (7 months lead time) and evaluates them at three study sites in the Italian Alps which are relevant for water management, hydropower production and alpine ski tourism. The prototype to generate snow forecasts is based on the SNOWPACK model forced by meteorological data of two Copernicus Climate Data Store seasonal forecast systems, namely ECMWFS5 and MFS6. Forecast skill has been assessed employing both deterministic and probabilistic metrics, and using snow depth station measurements as a reference. The skill has been investigated also in relation to different levels of post-processing of the total precipitation input, i.e. using raw, bias-corrected, downscaled, bias-corrected and downscaled precipitation data, since this variable deeply affects snow dynamics and the goodness of snow simulations.

Many robust features have been found across different seasonal forecast systems, seasons, stations and scores. The prototype running from November 1^{st} up to 7 months lead time, shows surprisingly good skill at predicting the tercile category for different time aggregation of the snow forecasts: below- and above-normal winter (DJF), spring (MAM), and November-May average snow depth are predicted with higher accuracy (BSS) and higher discrimination (AUCSS) with respect to a simple forecasting method based on the climatology. Ensemble mean monthly snow depth forecasts are significantly correlated with observations not only at short lead time 1 and 2 months (November and December) but also at lead time 5 and 6 months (March and April) when employing the ECMWFS5 forcing. Moreover the prototype shows skill at predicting extremely dry seasons, i.e. seasons with snow depth below the 10^{th} percentile, while the prediction of extremely wet seasons (i.e. snow depth above the 90^{th} percentile) is model-, station- and score-dependent. The bias adjustment of precipitation forecasts with the quantile mapping technique can substantially improve the agreement between the modelled and the observed snow depth climatology provided that a reliable reference dataset is used. However, the application of bias-adjustment, downscaling or bias-adjustment and downscaling techniques does not result in remarkable differences on the skill scores compared to the case in which raw precipitation data are employed. This suggests that the probabilistic skill scores are weakly sensitive to the treatment of the precipitation input. The use of raw precipitation data allows simplifying the modelling chain and boosting the production of snow forecasts at least at the three study sites considered. The exportability of these results to other study sites should be checked.

The predictability of the snowpack deviation with respect to normal conditions at lead times up to 7 months is the major result of this study and corroborates the hypothesis that snowpack is a natural "integrator" of the climatic conditions (conditions of the meteorological drivers) at the monthly/seasonal scale, so even if the forecasts of the drivers (air temperature, precipitation, etc ...) do not exactly match the observations at sub-monthly time scales, the differences may compensate over monthly/seasonal time scales and provide reasonable monthly/seasonal snowpack forecasts. This is an important step forward in the seasonal prediction of hydrological variables: while the skill in streamflow prediction is limited, the storage of water within the snowpack can be predicted also at long lead time. This is particularly relevant in mountain catchments where most of the run-off in spring is due to snow melt, and the forecasts of below- or above-normal snow depth have immediate applications in the management of water resources, hydropower production and ski resort management. A reliable seasonal forecasting system, e.g with a lead-

time up to 3–6 months, could bring an important improvement in the long-term optimization of the energy production, since the hydropower reservoir management heavily depends on the expected seasonal hydrological characteristics, e.g. the snowpack development.

Although this prototype has been conceived to respond to practical needs of end users and it has been applied in specific study areas where forecasts were meaningful to them, it is extremely flexible and it can be applied to any other mountain areas, provided that long-term temperature and snow depth time series are available for bias-correcting temperature forecasts and validating snow predictions, respectively.

In light of the exportability of this prototype to any mountain site, future work should be done to run this prototype at other sites of the Alps and beyond to further check its skill and to obtain a more complete picture of the snow forecasts for the season ahead along elevational transects or at the regional or even mountain range scale. These forecasts are particularly useful for all activities and sectors related to snow-hydrological fields, i.e. for example irrigation consortia, industry, ski resort, hydropower plant and water resource managers. In addition, they help estimating the amount of water made available by snowmelt, mainly at the head of Alpine catchments, since in summer it accounts for almost the total runoff. This knowledge can help to better address problems related to dearth of water in drought periods, which are expected to become more and more frequent in the future in the Alpine region.

Data availability. The datasets presented in this study can be obtained upon request to the corresponding author.

Author contributions. Original idea of the work: ST, JvH, ; Development of the modelling chain: ST, with help of JvH for bias-correction and downscaling tools; Run simulations: ST with help of GB; Data Analysis: GB; Writing first draft of the paper: ST and GB; Revision of the paper: JvH and all

Competing interests. The authors declare that no competing interests are present.

Acknowledgements. This work was performed in the framework of the MEDSCOPE (MEDiterranean Services Chain based On climate PrEdictions) ERA4CS project (grant agreement no. 690462) funded by the European Union. We acknowledge fruitful discussions with the following end-users: i) IREN S.p.A. on hydropower producers needs, ii) the Water Resources Department of the Metropolitan City of Turin (NW Italy) on water management needs, iii) Monterosa 2000 S.p.A. on ski-resort and artificial snowmaking needs.

References

Anandhi, A.: Uncertainties in downscaled relative humidity for a semi-arid region in India, Journal of earth system science, 120, 375–386, 2011.

Anghileri, D., Voisin, N., Castelletti, A., Pianosi, F., Nijssen, B., and Lettenmaier, D. P.: Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments, Water Resources Research, 52, 4209–4225, 2016.

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, Hydrology and Earth System Sciences, 22, 2057–2072, 2018.

Athanasiadis, P. J., Bellucci, A., Scaife, A. A., Hermanson, L., Materia, S., Sanna, A., Borrelli, A., MacLachlan, C., and Gualdi, S.: A multisystem view of wintertime NAO seasonal predictions, Journal of Climate, 30, 1461–1475, 2017.

Bartelt, P. and Lehning, M.: A physical SNOWPACK model for the Swiss avalanche warning: Part I: numerical model, Cold Regions Science and Technology, 35, 123–145, 2002.

Becker, B. D.: DRAFT Metaxa set: A new synthetic European windstorm event set, Tech. rep., Met Office, 2019.

Bradley, A. A., Schwartz, S. S., and Hashino, T.: Sampling uncertainty and confidence intervals for the Brier score and Brier skill score, Weather and Forecasting, 23, 992–1006, 2008.

BSC-CNS, Guemas, V., Manubens, N., Garcia-Serrano, J., Fuckar, N., Caron, L.-P., Bellprat, O., Rodrigues, L., Torralba, V., Hunter, A., Prodhomme, C., and Menegoz, M.: s2dverification: Set of Common Tools for Forecast Verification, https://CRAN.R-project.org/package=s2dverification, r package version 2.10.0, 2021.

Calì Quaglia, F., Terzago, S., and von Hardenberg, J.: Temperature and precipitation seasonal forecasts over the Mediterranean region: added value compared to simple forecasting methods, Climate Dynamics, pp. 1–25, 2021.

Dorel, L., Ardilouze, C., Déqué, M., Batté, L., and Guérémy, J.-F.: Documentation of the METEO-FRANCE Pre-Operational seasonal forecasting system, Service contract n° 2015/c3s_433_lot1-meteo-france (Deliverable D3.1), Météo-France, 2017.

D'Onofrio, D., Palazzi, E., von Hardenberg, J., Provenzale, A., and Calmanti, S.: Stochastic rainfall downscaling of climate models, Journal of Hydrometeorology, 15, 830–843, 2014.

Fick, S. E. and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, International journal of climatology, 37, 4302–4315, 2017.

Förster, K., Hanzer, F., Stoll, E., Scaife, A. A., MacLachlan, C., Schöber, J., Huttenlau, M., Achleitner, S., and Strasser, U.: Retrospective forecasts of the upcoming winter season snow accumulation in the Inn headwaters (European Alps), Hydrology and Earth System Sciences, 22, 1157–1173, https://doi.org/10.5194/hess-22-1157-2018, 2018.

Greuell, W., Franssen, W. H., Biemans, H., and Hutjes, R. W.: Seasonal streamflow forecasts for Europe–Part I: Hindcast verification with pseudo-and real observations, Hydrology and Earth System Sciences, 22, 3453–3472, 2018.

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., and Engen-Skaugen, T.: Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations–a comparison of methods, Hydrology and Earth System Sciences, 16, 3383–3390, 2012.

Gupta, A. S. and Tarboton, D. G.: A tool for downscaling weather data from large-grid reanalysis products to finer spatial scales for distributed hydrological applications, Environmental Modelling & Software, 84, 50–69, 2016.

Haslinger, K., Koffler, D., Schöner, W., and Laaha, G.: Exploring the link between meteorological drought and streamflow: Effects of climate-catchment interaction, Water Resources Research, 50, 2468–2487, 2014.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Hirashima, H., Yamaguchi, S., Sato, A., and Lehning, M.: Numerical modeling of liquid water movement through layered snow based on new measurements of the water retention curve, Cold Regions Science and Technology, 64, 94–103, 2010.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P. E., Mogensen, K., Zuo, H., and Monge-Sanz, B. M.: SEAS5: the new ECMWF seasonal forecast system, Geoscientific Model Development, 12, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019, 2019.

Jolliffe, I. T. and Stephenson, D. B.: Forecast verification: a practitioner's guide in atmospheric science, John Wiley & Sons, 2012.

Kapnick, S. B., Yang, X., Vecchi, G. A., Delworth, T. L., Gudgel, R., Malyshev, S., Milly, P. C., Shevliakova, E., Underwood, S., and Margulis, S. A.: Potential for western US seasonal snowpack prediction, Proceedings of the National Academy of Sciences, 115, 1180–1185, 2018.

Kochendorfer, J., Nitu, R., Wolff, M., Mekis, E., Rasmussen, R., Baker, B., Earle, M. E., Reverdin, A., Wong, K., Smith, C. D., et al.: Analysis of single-Alter-shielded and unshielded measurements of mixed and solid precipitation from WMO-SPICE, Hydrology and Earth System Sciences, 21, 3525–3542, 2017a.

Kochendorfer, J., Rasmussen, R., Wolff, M., Baker, B., Hall, M. E., Meyers, T., Landolt, S., Jachcik, A., Isaksen, K., Brækkan, R., et al.: The quantification and correction of wind-induced precipitation measurement errors, Hydrology and Earth System Sciences, 21, 1973–1989, 2017b.

Köberl, J., François, H., Cognard, J., Carmagnola, C., Prettenthaler, F., Damm, A., and Morin, S.: The demand side of climate services for real-time snow management in Alpine ski resorts: Some empirical insights and implications for climate services development, Climate Services, 22, 100 238, https://doi.org/https://doi.org/10.1016/j.cliser.2021.100238, 2021.

Lawrence, M. G.: The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications, Bulletin of the American Meteorological Society, 86, 225–234, 2005.

Lehning, M., Bartelt, P., Brown, B., and Fierz, C.: A physical SNOWPACK model for the Swiss avalanche warning: Part III: Meteorological forcing, thin layer formation and evaluation, Cold Regions Science and Technology, 35, 169–184, 2002.

Li, D., Lettenmaier, D. P., Margulis, S. A., and Andreadis, K.: The value of accurate high-resolution and spatially continuous snow information to streamflow forecasts, Journal of Hydrometeorology, 20, 731–749, 2019.

Lledó, L., Cionni, I., Torralba, V., Bretonnière, P.-A., and Samsó, M.: Seasonal prediction of Euro-Atlantic teleconnections from multiple systems, Environmental Research Letters, 15, 074 009, 2020.

Mair, E., Bertoldi, G., Leitinger, G., Della Chiesa, S., Niedrist, G., and Tappeiner, U.: ESOLIP–estimate of solid and liquid precipitation at sub-daily time resolution by combining snow height and rain gauge measurements, Hydrology and Earth System Sciences Discussions, 10, 8683–8714, 2013.

Marke, T., Strasser, U., Hanzer, F., Stötter, J., Wilcke, R. A. I., and Gobiet, A.: Scenarios of future snow conditions in Styria (Austrian Alps), Journal of Hydrometeorology, 16, 261–277, 2015.

Martinez-García, F. P., Contreras-de Villar, A., and Muñoz-Perez, J. J.: Review of Wind Models at a Local Scale: Advantages and Disadvantages, Journal of Marine Science and Engineering, 9, 318, 2021.

Mason, S.: Guidance on verification of operational seasonal climate forecasts. WMO 1220, 81 pp, 2018.

Mason, S. J.: On Using "Climatology" as a Reference Strategy in the Brier and Ranked Probability Skill Scores, Monthly Weather Review, 132, 1891 – 1895, https://doi.org/10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2, 2004.

Matheson, J. E. and Winkler, R. L.: Scoring rules for continuous probability distributions, Management science, 22, 1087–1096, 1976.

Mengelkamp, H.-T., Kapitza, H., and Pflüger, U.: Statistical-dynamical downscaling of wind climatologies, Journal of Wind Engineering and Industrial Aerodynamics, 67, 449–457, 1997.

Mishra, N., Prodhomme, C., and Guemas, V.: Multi-model skill assessment of seasonal temperature and precipitation forecasts over Europe, Climate Dynamics, 52, 4207–4225, 2019.

Palazzi, E., Mortarini, L., Terzago, S., and Von Hardenberg, J.: Elevation-dependent warming in global climate model simulations at high spatial resolution, Climate Dynamics, 52, 2685–2702, 2019.

Pepin, N., Bradley, R. S., Diaz, H., Baraër, M., Caceres, E., Forsythe, N., Fowler, H., Greenwood, G., Hashmi, M., Liu, X., et al.: Elevation-dependent warming in mountain regions of the world, Nature climate change, 5, 424–430, 2015.

Perez-Zanon, N., Caron, L.-P., Alvarez-Castro, C., Batte, L., von Hardenberg, J., LLedo, L., Manubens, N., Sanchez-Garcia, E., van Schaeybroeck, B., Torralba, V., and Verfaillie, D.: CSTools: Assessing Skill of Climate Forecasts on Seasonal-to-Decadal Timescales, https://CRAN.R-project.org/package=CSTools, r package version 4.0.0, 2021.

Pörtner, H.-O., Roberts, D. C., Masson-Delmotte, V., Zhai, P., Tignor, M., Poloczanska, E., and Weyer, N.: The ocean and cryosphere in a changing climate, 2019.

Pryor, S. and Barthelmie, R.: Hybrid downscaling of wind climates over the eastern USA, Environmental Research Letters, 9, 024 013, 2014.

Pryor, S. and Hahmann, A. N.: Downscaling wind, in: Oxford Research Encyclopedia of Climate Science, 2019.

Rebora, N., Ferraris, L., von Hardenberg, J., and Provenzale, A.: RainFARM: Rainfall downscaling by a filtered autoregressive model, Journal of Hydrometeorology, 7, 724–738, 2006.

Salameh, T., Drobinski, P., Vrac, M., and Naveau, P.: Statistical downscaling of near-surface wind over complex terrain in southern France, Meteorology and Atmospheric Physics, 103, 253–265, 2009.

Santos, I. M., Herrnegger, M., and Holzmann, H.: Seasonal discharge forecasting for the Upper Danube, Journal of Hydrology: Regional Studies, 37, 100 905, 2021.

Schulzweida, U.: CDO User Guide, Tech. Rep. October, MPI for Meteorology, 2019.

Stahl, K., Kohn, I., Blauhut, V., Urquijo, J., De Stefano, L., Acácio, V., Dias, S., Stagge, J. H., Tallaksen, L. M., Kampragou, E., et al.: Impacts of European drought events: insights from an international database of text-based reports, Natural Hazards and Earth System Sciences, 16, 801–819, 2016.

Stephan, R., Erfurt, M., Terzi, S., Žun, M., Kristan, B., Haslinger, K., and Stahl, K.: An Alpine Drought Impact Inventory to explore past droughts in a mountain region, Natural Hazards and Earth System Sciences Discussions, pp. 1–25, 2021.

Terzago, S., Palazzi, E., and Hardenberg, J. v.: Stochastic downscaling of precipitation in complex orography: A simple method to reproduce a realistic fine-scale climatology, Natural Hazards and Earth System Sciences, 18, 2825–2840, 2018.

Terzago, S., Andreoli, V., Arduini, G., Balsamo, G., Campo, L., Cassardo, C., Cremonese, E., Dolia, D., Gabellani, S., von Hardenberg, J., et al.: Sensitivity of snow models to the accuracy of meteorological forcings in mountain environments, Hydrology and Earth System Sciences, 24, 4061–4090, 2020.

Wanders, N., Thober, S., Kumar, R., Pan, M., Sheffield, J., Samaniego, L., and Wood, E. F.: Development and evaluation of a pan-European multimodel seasonal hydrological forecasting system, Journal of Hydrometeorology, 20, 99–115, 2019.

Wever, N., Schmid, L., Heilig, A., Eisen, O., Fierz, C., and Lehning, M.: Verification of the multi-layer SNOWPACK model with different water transport schemes, The Cryosphere, 9, 2271–2293, https://doi.org/10.5194/tc-9-2271-2015, 2015.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.