Using a 10-Year Radar Archive for Nowcasting Precipitation Growth and Decay: A Probabilistic Machine Learning Approach

LORIS FORESTI AND IOANNIS V. SIDERIS

Federal Office of Meteorology and Climatology, MeteoSwiss, Locarno-Monti, Switzerland

DANIELE NERINI

Federal Office of Meteorology and Climatology, MeteoSwiss, Locarno-Monti, and Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

LEA BEUSCH

Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

URS GERMANN

Federal Office of Meteorology and Climatology MeteoSwiss, Locarno-Monti, Switzerland

(Manuscript received 18 December 2018, in final form 25 June 2019)

ABSTRACT

Machine learning algorithms are trained on a 10-yr archive of composite weather radar images in the Swiss Alps to nowcast precipitation growth and decay in the next few hours in moving coordinates (Lagrangian frame). The hypothesis of this study is that growth and decay is more predictable in mountainous regions, which represent a potential source of practical predictability by machine learning methods. In this paper, artificial neural networks (ANN) are employed to learn the complex nonlinear dependence relating the growth and decay to the input predictors, which are geographical location, mesoscale motion vectors, freezing level height, and time of the day. The average long-term growth and decay patterns are effectively reproduced by the ANN, which allows exploring their climatology for any combination of predictors. Due to the low intrinsic predictability of growth and decay, its prediction in real time is more challenging, but is substantially improved when adding persistence information to the predictors, more precisely the growth and decay and precipitation intensity in the immediate past. The improvement is considerable in mountainous regions, where, depending on flow direction, the root-mean-square error of ANN predictions can be 20%–30% lower compared with persistence. Because large uncertainty is associated with precipitation forecasting, deterministic machine learning predictions should be coupled with a model for the predictive uncertainty. Therefore, we consider a probabilistic perspective by estimating prediction intervals based on a combination of quantile decision trees and ANNs. The probabilistic framework is an attempt to address the problem of conditional bias, which often characterizes deterministic machine learning predictions obtained by error minimization.

1. Introduction

Forecasting precipitation in the very short range (0–2 h) commonly relies on extrapolation-based nowcasting tools that exploit the persistence of the most recent weather radar observations (see e.g., Germann and Zawadzki 2002). In this time range, many critical decisions are taken to ensure people's safety (e.g., closing of train lines susceptible to debris flow, optimization of airport operations, and evacuation of vulnerable construction zones; see e.g., Germann et al. 2017). Because the costs related to such interruptions are high, these activities need to remain

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/WAF-D-18-0206s1>.

Corresponding author: Loris Foresti, loris.foresti@meteoswiss.ch

DOI: 10.1175/WAF-D-18-0206.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

FIG. 1. Growth and decay occurring when precipitation moves over orography. The goal of this paper is to use machine learning to predict the growth and decay in moving coordinates (Lagrangian frame).

operational despite the warnings of severe weather issued by forecasters one or more days ahead. A short interruption is considered only when the probability of occurrence or potential damage of a localized hazard is very high. To obtain the best possible prediction skill in the 0–2-h range, one cannot solely rely on numerical weather prediction (NWP) but must also use the available observations in a more direct way. Therefore, improvements of extrapolation-based nowcasting tools can have an important impact on these activities.

a. Sources of uncertainty in persistence-based nowcasting of radar precipitation fields

Sequences of radar precipitation fields exhibit persistence in both the Eulerian and the Lagrangian frame, where the latter assumes persistence in the coordinates moving with the storm (e.g., Zawadzki 1973; Germann and Zawadzki 2002). The forecasting procedure based on Lagrangian persistence involves using an optical flow method to estimate a field of radar echo motion and applying an advection scheme to produce an extrapolation nowcast (e.g., Germann and Zawadzki 2002).

The uncertainty of radar-based extrapolation nowcasts can be categorized into the following main classes (adapted from Germann et al. 2006):

- 1) *Initial condition uncertainty* related to *radar measurement* errors. In a well maintained and calibrated radar network, the main error sources are related to the space–time variability of the vertical profile of reflectivity (VPR) and the Z – R relationship, partial and total beam blockage, and signal attenuation (e.g., Villarini and Krajewski 2010). The uncertainty also includes spatial and temporal sampling errors of the radar measurements, which may affect the estimation of the radar echo motion (see point 2).
- 2) *Model uncertainty* related to imperfections of the nowcasting model and of the selection of *model parameters*. The main errors are due to inaccuracies of the algorithm for the motion field retrieval, the choice of model parameters and, to a lesser degree, the numerical diffusion of the advection scheme.

3) *Model uncertainty* related to the *assumption of persistence* of the atmospheric state. This comprises the unknown future evolution of

- (i) the precipitation intensity (i.e., its initiation, growth, decay, and termination),
- (ii) the motion field, and
- (iii) the statistical properties of precipitation fields (e.g., spatial and temporal autocorrelations, Fourier spectra, degree of intermittency, and probability density function).

The uncertainty of radar-based quantitative precipitation estimation (QPE; point 1 above) can be an important source of nowcast errors in the first hour (e.g., Fabry and Seed 2009). A common approach to characterize this uncertainty is to generate QPE ensembles (e.g., Germann et al. 2009).

For a well-designed nowcasting system starting from a good quality radar QPE product, the main source of uncertainty beyond a lead time of about 30 min arises from precipitation initiation, growth, decay, and termination processes that violate the persistence assumption [point 3(i) above, Bowler et al. 2006; Germann et al. 2006].

The focus of our study is on the predictability of *growth and decay* (GD), which is the first time derivative of a radar precipitation time series in the Lagrangian frame (see Fig. 1). Precipitation initiation and termination processes are not considered in this study.

b. Predictability of precipitation growth and decay

In this paper, the term predictability refers to the *practical predictability* of the atmosphere, defined as the extent to which a forecasting technique, in our case, an extrapolation-based or machine learning-based method, provides useful prediction skill (see e.g., Lorenz 1996; Surcel et al. 2015).

The precipitation GD can be decomposed into a *predictable and unpredictable component*. Most ensemble nowcasting systems do not attempt to predict the GD trend (i.e., the predictable part), but only generate stochastic ensemble members as a way to estimate the forecast uncertainty. Examples of nowcasting systems exploiting the latter

strategy are the Short-Term Ensemble Prediction System (STEPS; Bowler et al. 2006; Seed et al. 2013), the String of Beads Model for Nowcasting (SBMcast; Berenguer et al. 2011) and the stochastic extension of the McGill Algorithm for Precipitation Nowcasting by Lagrangian Extrapolation (MAPLE; Atencia and Zawadzki 2014).

Radhakrishna et al. (2012) studied the scale dependence of the predictability of growth and decay fields by Lagrangian persistence using data from the U.S. national radar composite. Results show that precipitation fields are much more persistent than GD fields, which explains in part why previous attempts of predicting the trend of thunderstorm intensity did not significantly improve the forecast skill (e.g., Tsonis and Austin 1981; Wilson et al. 1998). More precisely, Radhakrishna et al. (2012) found that GD patterns are persistent up to a lead time of 2 h but only for scales larger than 250 km over the continental United States.

The hypothesis underpinning our study is that precipitation GD is more predictable in mountainous regions, which represent a potential source of practical predictability. Compared to the flat continental United States, the predictable spatial scales are expected to be smaller over orography (see e.g., Foresti and Seed 2015; Foresti et al. 2018).

c. Machine learning applications in weather forecasting

First models for statistical weather prediction appeared in the 1950s (e.g., Malone 1955; Lorenz 1956). Machine learning (ML) deals with similar statistical tasks (e.g., classification and regression), but focuses on designing flexible algorithms that maximize predictive power (Breiman 2001b). Statistical weather forecasting with machine learning started in the early 1990s (e.g., McCann 1992; Kuligowski and Barros 1998; Hall et al. 1999) and became more widespread after the year 2000 (see reviews by Haupt et al. 2009; McGovern et al. 2017). The domains of application include, among others, the processing of remote sensing observations (e.g., Marzban and Witt 2001; Foresti et al. 2012; Besic et al. 2016; Beusch et al. 2018), NWP postprocessing (e.g., Kretschmar et al. 2004; Taillardat et al. 2016; Gagne et al. 2017; Rasp and Lerch 2018), nowcasting and short-range forecasting (e.g., Manzato 2005; Mecikalski et al. 2015; Han et al. 2017; Sprenger et al. 2017; Ukkonen et al. 2017).

Machine learning surged in popularity in recent years thanks to various advances in computer hardware and algorithms. Processing of large datasets was made possible by the increase in computer memory, storage, and network capabilities. A notable example is the graphics processing unit (GPU) technology, which allows training deeper and more complex artificial neural network

(ANN) architectures (e.g., Fukushima and Miyake 1982; Hinton et al. 2006; Goodfellow et al. 2016). In parallel, better training can be obtained by stochastic optimization routines (e.g., Kingma and Ba 2015), while the vanishing gradient problem can be mitigated by using different activation functions [e.g., the rectified linear unit (ReLU); Glorot et al. 2011]. For a historical overview on deep learning, we refer to Schmidhuber (2015).

To our knowledge, the first study that tested the usage of ANNs for precipitation nowcasting is by French et al. (1992). The authors trained an ANN to predict the evolution of synthetic rainfall fields, but did not find significantly higher skill compared to Lagrangian persistence.

Grecu and Krajewski (2000) went a step further by separating the prediction problem into two steps: the estimation of the radar echo motion and the use of ANN for statistical prediction of the dynamic precipitation changes (GD). Using radar data from Tulsa, Oklahoma, they did not find a substantial improvement compared to Lagrangian persistence either.

Given the increasing size of radar data archives (Carbone et al. 2002; Fabry et al. 2017; Peleg et al. 2018), it becomes now possible to study the dependence of the predictability of GD on spatial location, time of day, orography and flow conditions. In short, there is potential to better understand, predict and correct the forecast error of persistence-based nowcasts.

d. Objectives of this study

The aim of this study is to use machine learning to bring precipitation nowcasting beyond the assumption of Lagrangian persistence. The current paper completes the work of Foresti et al. (2018), who used the same 10-yr archive of radar composite fields in the Swiss Alpine region to derive a climatology of precipitation GD depending on geographical location, freezing level height, mesoscale flow direction, and speed (input predictors).

The first goal of this study is to use ML, more precisely artificial neural networks, to automatically learn the localized dependence of precipitation GD on the input predictors.

The second goal is to estimate the relative importance of input predictors, and evaluate whether the machine learning nowcasts of GD can outperform a reference model based on persistence.

The third goal is to extend ML to give an indication of the forecast uncertainty. This is achieved by computing prediction intervals using a combination of ANN and decision trees (DT).

In this paper, we do not attempt to achieve the best possible predictive performance using the most advanced machine learning methods, but instead we aim to better understand the implications of the machine learning

FIG. 2. Map of the study domain and the location of weather radars (LEM: Lema, ALB: Albis, DOL: Dôle, PPM: Plaine Morte, WEI: Weissfluhgipfel). The radars covering the dataset period are displayed in red (LEM, ALB, DOL). Two example precipitation boxes at the origin and destination are also shown.

approach as a whole. In particular, we focus on the consequences of error minimization and the importance of uncertainty quantification in the context of weather forecasting.

e. Outline of the paper

The paper is structured as follows. [Section 2](#) formulates the statistical learning framework for nowcasting. [Section 3](#) describes the precipitation growth and decay dataset. [Section 4](#) briefly reviews the used machine learning algorithms. [Section 5](#) illustrates the prediction results and their verification. [Section 6](#) introduces the probabilistic machine learning framework and a new method to quantify the prediction uncertainty. Probabilistic GD predictions are shown and verified in [section 7](#). Finally, [sections 8](#) and [9](#) put the contributions into perspective and conclude the paper.

2. Statistical nowcasting frameworks

a. Nowcasting by persistence

Nowcasting by Lagrangian persistence can be formulated starting from the two-dimensional conservation equation by neglecting the compressibility term ([Germann and Zawadzki 2002](#)):

$$\frac{dR}{dt} = \frac{\partial R}{\partial t} + u \frac{\partial R}{\partial x} + v \frac{\partial R}{\partial y}, \quad (1)$$

where R is the rainfall intensity (or radar reflectivity), dR/dt is the source/sink term (growth and decay), $\partial R/\partial t$

is the local rate of change, and (u, v) is the flow vector. If we set the source/sink term to zero, we can derive a Lagrangian persistence nowcast by advecting the rainfall R along the motion field (see e.g., [Germann and Zawadzki 2002](#)):

$$R(t + \tau, \mathbf{s}) = R(t, \mathbf{s} - \boldsymbol{\alpha}), \quad (2)$$

where τ is the forecast lead time, $\mathbf{s} = (X, Y)$ is the spatial location (coordinates), and $\boldsymbol{\alpha} = (u, v) \cdot \tau$ is the displacement vector. In presence of GD ($dR/dt \neq 0$), (2) becomes

$$R(t + \tau, \mathbf{s}) = R(t, \mathbf{s} - \boldsymbol{\alpha}) + \text{GD}, \quad (3)$$

where GD is the growth and decay term in moving coordinates (see [Figs. 1](#) and [2](#)).

b. Nowcasting growth and decay with machine learning

Instead of relying only on a short sequence of radar fields and assuming persistence of GD ([Radhakrishna et al. 2012](#)), a machine learning approach potentially allows recurrent patterns to be learned from historical archives to be then applied for prediction.

Let us define y as the target variable that we seek to predict (i.e., $y = \text{GD}$). The statistical prediction of GD involves the estimation of a function f as follows:

$$y(t + \tau, \mathbf{s}) = f[y(t, \mathbf{s} - \boldsymbol{\alpha}_y), \mathbf{x}(t, \mathbf{s} - \boldsymbol{\alpha}_x)] + \varepsilon, \quad (4)$$

where $y(t, \mathbf{s} - \boldsymbol{\alpha}_y)$ is the current GD value, $\mathbf{x}(t, \mathbf{s} - \boldsymbol{\alpha}_x)$ is a vector of external predictors, and ε is the noise term (i.e., the unpredictable part). If $\boldsymbol{\alpha}_y \neq \mathbf{0}$ the values are retrieved upstream (Lagrangian frame), while if $\boldsymbol{\alpha}_y = \mathbf{0}$ they are retrieved at the same location (Eulerian frame); the same can be applied to $\boldsymbol{\alpha}_x$. Note that one can also use a sequence of k previous y and $\mathbf{x}$ values, as well as their temporal increments [e.g., $\mathbf{x}' = \mathbf{x}(t, \mathbf{s}) - \mathbf{x}(t - 1, \mathbf{s})$].

The sequence of previous GD values exploits persistence and represents an *endogenous* variable, while the external predictors are *exogenous* variables. Foresti et al. (2018) provides a comprehensive analysis of the dependence of GD on external predictors over the Swiss Alps, such as the freezing level height, the flow direction, and the geographical location.

In this study, we will train ANNs to predict the GD of the next hour ($\tau = 1$ h) and perform the following experiments:

- 1) $y(t + \tau, \mathbf{s}) = y(t, \mathbf{s})$,
- 2) $y(t + \tau, \mathbf{s}) = f[y(t, \mathbf{s}), \mathbf{s}]$,
- 3) $y(t + \tau, \mathbf{s}) = f[\mathbf{x}(t, \mathbf{s} - \boldsymbol{\alpha}), \mathbf{s}]$, and
- 4) $y(t + \tau, \mathbf{s}) = f[y(t, \mathbf{s}), \mathbf{x}(t, \mathbf{s} - \boldsymbol{\alpha}), \mathbf{s}]$.

In summary, experiment 1 assumes Eulerian persistence of the GD, experiment 2 uses ML to improve the persistence nowcast based on the archive and the spatial location, experiment 3 uses ML to predict the GD using only the set of exogenous predictors, and experiment 4 uses ML to predict the GD using both exogenous and endogenous predictors.

Experiment 3 is actually a reformulation of the stratified climatology of Foresti et al. (2018) as a supervised statistical learning problem. In fact, the conditional expectation of GD for a subset of weather conditions characterized by the predictors $\mathbf{x}$ is given by

$$E(y|\mathbf{x} \in \mathbb{X}_\omega) = \frac{1}{N} \sum_i y_i \in \mathbb{Y}_\omega$$

$$\lim_{\omega \rightarrow 0} E(y|\mathbf{x} \in \mathbb{X}_\omega) = f(y|\mathbf{x}) = f(\mathbf{x}) + \varepsilon, \quad (5)$$

where ω is the width of the stratification interval, $\mathbb{X}_\omega$ is the set of predictor values that fall within that interval, and $\mathbb{Y}_\omega$ is the corresponding set of GD values. As will be shown later, it is also possible to compute various moments of the distribution of y . In other words, ML offers the opportunity to estimate the marginal statistics of a variable y for an infinitesimally small interval width.

3. The radar precipitation growth and decay dataset

The radar archive covers the 10-yr period 2005–14 and comprises data from the Swiss C-band radars located at

Monte Lema, Albis, and La Dôle, which were completely renewed and upgraded to dual-polarization in 2011 and 2012. The radar network was extended with 2 new radars in 2014 and 2016, respectively (see Germann et al. 2017). Composite radar images have a spatial resolution of 1 km and a temporal resolution of 5 min.

The preparation of the precipitation GD dataset is described in Foresti et al. (2018). In summary, the procedure involves the following steps:

- 1) Estimating fields of radar echo motion using the MAPLE variational echo tracking (Germann and Zawadzki 2002).
- 2) Calculating backward trajectories of radar echoes.
- 3) Defining a regular grid of overlapping boxes of a given size.
- 4) Computing the *mean areal precipitation* (MAP; mm h^{-1}) for each box using the destination location at time $t = t + \tau$ and the origin located upstream at time t following the trajectories.

Note that the methods of points 1 and 2 are available in the open-source python library “pysteps” (Pulkkinen et al. 2019, manuscript submitted to *Geosci. Model Dev. Discuss.*). As in Foresti et al. (2018), we selected a lead time of $\tau = 1$ h and a box size of $64 \times 64 \text{ km}^2$, which are regularly distributed on an 8-km resolution grid. Throughout the paper, all the machine learning predictions and verification are done at these spatial and temporal resolutions. Note that in addition to the MAP, one can also derive other precipitation statistics, such as the fraction of wet pixels.

As explained in section 2, the quantity of interest for nowcasting is the precipitation GD, which is defined here as the multiplicative difference between the MAP at the origin and destination location (Foresti et al. 2018):

$$\text{GD} = y = 10 \log_{10} \left(\frac{\text{MAP}_d + c}{\text{MAP}_o + c} \right) (\text{dB}), \quad (6)$$

where $c = 0.01$ is a small constant offset to avoid the division by 0, $\text{MAP}_d = \text{MAP}(t + 1, \mathbf{s})$ is the MAP of the destination box at time $t + 1$ and location $\mathbf{s}$ and $\text{MAP}_o = \text{MAP}(t, \mathbf{s} - \boldsymbol{\alpha})$ is the MAP of the origin box located upstream following the displacement vector $\boldsymbol{\alpha}$ at time t (see Fig. 2). A backward-in-time semi-Lagrangian scheme is used to advect the origin box to the destination by using all the 5-min motion fields. This procedure helps to isolate the GD error from the one due to the nonstationary motion (Germann et al. 2006).

The multiplicative formulation of GD has two advantages: 1) it makes the distribution of GD symmetric around 0 and close to normal, and 2) it reduces the correction of a persistence nowcast to a summation:

TABLE 1. Structure of data archive for training the machine learning algorithms. On the left is the set of input predictors and on the right the output predictand(s). In a real-time application the destination location $[X_d, Y_d]$ is found by assuming stationarity of the motion vectors $[U, V]$ during the nowcast.

Input predictors								Output predictand(s)	
Exogenous predictors				Endogenous predictors				Target variable(s)	
X_d	Y_d	U	V	$\text{HZT}_o(t)$	$D_{\sin}(t+1)$	$D_{\cos}(t+1)$	$\text{MAP}_o(t)$	$\text{GD}_d(t)$	
Location		Flow		Airmass	Daytime		Precipitation	Growth/decay	
								$\text{MAP}_d(t+1)$	$\text{GD}_d(t+1)$
								Precipitation	Growth/decay

$$10 \log_{10}(\text{MAP}_d^{\text{pred}}) = 10 \log_{10}(\text{MAP}_o) + \text{GD}_{\text{pred}}(\text{dB}). \quad (7)$$

The radar archive was extended to contain the freezing level height (HZT), which was extracted from the hourly analyses of the COSMO NWP model (Baldauf et al. 2011). In fact, Foresti et al. (2018) found that the spatial distribution of GD depends on HZT, which constitutes a useful proxy of the air stability.

An additional forcing variable that could potentially control the GD is the diurnal cycle (see e.g., Mandapaka et al. 2013; Atencia et al. 2017; Fabry et al. 2017). The time of the day, $h \in \mathbb{R}: 0 \leq h < 24$, is a circular variable. To ensure the continuity of the predictor at midnight, the hour of the day h was coded using the following two variables:

$$D_{\sin} = \sin(2\pi h/24), \quad D_{\cos} = \cos(2\pi h/24). \quad (8)$$

The season could be characterized in a similar way. However, it was found that seasonal variability is better explained by the freezing level height (Foresti et al. 2018).

The structure of the data archive is presented in Table 1. The main target variable is the GD term, although it is also possible to directly predict the $\text{MAP}_d(t+1)$. We decided to classify MAP_o as an endogenous predictor since it contributes to the definition of GD [see (6)]. Note that the predictor $\text{GD}_d(t)$ is in Eulerian coordinates (i.e., it is at the same spatial location). In fact, in the Alpine region the orographic forcing generates precipitation, and consequently GD patterns, that can remain persistent on the same location for several hours (e.g., Panziera et al. 2011). Therefore, we did not perform experiments using the Lagrangian GD [i.e., $\text{GD}_o(t)$].

Over the 10-yr period, we collected more than 21 million boxes (samples) with precipitation at both the origin and destination. Cases of precipitation initiation and termination are discarded to simplify the learning problem. In fact, the choice of predictors was not targeted for nowcasting the initiation of convective cells.

Finally, it is important to mention that, despite using a high-quality radar rainfall product, a certain fraction of

GD is related to the variability of radar coverage over the Swiss Alps. This was the main motivation for the installation of the two new radars in the Valais (PPM) and the Grisons (WEI). For a more detailed discussion on radar data uncertainty and GD, we refer to Foresti et al. (2018).

4. Machine learning algorithms and training

Supervised machine learning provides flexible algorithmic tools to solve tasks such as robust nonlinear classification and regression of data in high-dimensional spaces (Haykin 1998; Breiman 2001a; Goodfellow et al. 2016).

Compared with traditional statistical data models, the algorithmic approach is fully nonparametric, that is, it does not require making strong assumptions about the data distribution (e.g., Gaussianity) or the form of statistical dependency between variables (e.g., linearity). Instead, it is designed to maximize prediction skill while being robust to the curse of dimensionality (see e.g., Breiman 2001b).

a. Artificial neural networks

In this study, we used a feedforward artificial neural network model known as multilayer perceptron (MLP). The MLP architecture is composed of one input layer, one or more hidden layers, and one output layer, whose neurons are connected by synaptic weights (see an example in Fig. 3). The number of neurons in the input layer is equal to the number of input predictors. The output layer usually contains one single neuron with the target variable to predict (predictand). Alternatively, a multioutput MLP can be designed for joint prediction of multiple target variables, as will be explained in section 6b. The hidden layer(s) contain a set of neurons performing a nonlinear transformation (activation) of the weighted linear summation of values coming from the input neurons.

The MLP training consists of an iterative optimization of the network weights to minimize the error between the predicted and the target values in the output neuron. In this study, we used the mean square error (MSE, L2-norm). Given the nonconvexity of the error function

FIG. 3. Example of single-output MLP to predict the mean growth and decay using a set of external predictors.

and presence of multiple local minima, it is advised to use stochastic gradient descent optimization algorithms.

b. Tree-based methods

Classification and regression trees were introduced by Breiman et al. (1984). Their success stems mainly from their conceptual simplicity, which mirrors human decision-making and simplifies the interpretation of results. Also, they are very efficient with large datasets and can handle both numerical and categorical data without the need for data preprocessing.

The learning phase of decision trees involves a recursive partitioning of the training set into a set of leaves to minimize a given error function. As we perform regression, in this paper we used the MSE.

Decision-tree learning is conceptually similar to the process of data stratification presented in Foresti et al. (2018) and at the end of section 2b. The difference is that in the data stratification the splitting is done at regular intervals on the predictors (flow direction, HZT, etc.), with the exception of the spatial coordinates, where no splitting is performed.

The generalization error of decision trees can be improved by averaging the results of an ensemble of trees (e.g., as done by random forests; Breiman 2001a) or by boosting the prediction of an ensemble of “weak” trees (e.g., as done by the AdaBoost algorithm; Freund and Schapire 1997).

c. Data splitting and hyperparameter selection

Following good practices in machine learning (e.g., Marzban and Witt 2001; Kanevski et al. 2009), we randomly split the precipitation events into three sets: one

for training (60% of samples), one for validation (20%), and one for testing purposes (20%). The random splitting of precipitation events, instead of individual radar boxes, is essential to remove serial correlation between the different sets. This allows for more realistic estimations of the generalization error by preventing overfitting.

The training set is used to train the model, for example to find the optimal ANN weights. The validation set is used to tune the model hyperparameters and control the complexity of the function (to avoid over and underfitting) (e.g., by varying the number of hidden neurons in the ANN). Finally, the test set is used to estimate the generalization error, as the one derived from the validation set is slightly optimistic.

The experiments were carried out using the Python library “scikit-learn” (sklearn; Pedregosa et al. 2011). The input predictors were scaled to zero mean and unit variance. For the ANN, we selected a sufficiently large number of hidden neurons and applied an early stopping procedure, which saves the network weights that minimize the validation error during training. Different hidden layer sizes were tested in combination with the early stopping procedure and we finally selected one hidden layer with 100 neurons for the experiments. The ANN was trained using an initial learning rate of 0.001 and the stochastic gradient descent algorithm Adam (Kingma and Ba 2015), which exploits the first and second moment of the gradient to adapt the learning rate of individual parameters. In addition to computational speed, Adam is expected to work well with precipitation data, which have a substantial amount of stochastic variability and noise. Default values were kept for all other parameters. We also verified that the validation error reached a minimum before the maximum number of iterations and used the occurrence of overfitting as evidence that the network complexity is sufficient (e.g., Kanevski et al. 2009).

The decision tree hyperparameters were optimized by grid-search. The maximum tree depth was varied in the range [5, 10, 15, 20] and the minimum number of samples per leaf in the range [50, 100, 150, 200, 250]. The parameter combination that minimizes the validation error is selected. Default values were kept for all other parameters.

d. Bias-variance dilemma and intercomparison of forecast systems

Following the decomposition of the MSE into bias and variance components (i.e., $MSE = bias^2 + var$), one can see that the minimization of the MSE also minimizes the variance of the errors, which indirectly leads to minimizing the variance of the predictions. This has

FIG. 4. ANN predictions of the mean GD with different flow directions and freezing level heights for a fixed flow speed of 30 km h^{-1} . NW flows with (a) HZT at 1500 m and (b) HZT at 4000 m. SW flows with (c) HZT at 1500 m and (d) HZT at 4000 m.

practical consequences for the verification of the ANN, and the comparison with forecast systems having a larger variance (e.g., an extrapolation nowcast), which preserves the variance of the observations. Therefore, one should be cautious when comparing extrapolation-based with machine learning-based nowcasts.

The mentioned issues could be overcome either by normalizing the MSE or by comparing the two systems at the same spatial frequencies (e.g., by low-pass filtering the persistence nowcast; Seed 2003; Turner et al. 2004). The latter, however, is not directly applicable to our problem because of the intermittency of GD fields, which arises from the conditionality criterion (precipitation at both origin and destination).

Therefore, in this paper we choose to compare the root-mean-square error (RMSE) to the Pearson correlation coefficient (PCORR) and the regression slope β to gain further insight into how the bias-variance trade-off manifests itself in our problem. In our case, β measures the degree of *type 2 conditional bias*, that is, with respect to observations (Murphy 1995; Potts 2012):

$$\beta = \frac{\sigma_{\text{pred}}}{\sigma_{\text{obs}}} \rho, \quad (9)$$

where ρ is the PCORR between predictions and observations, and σ_{pred} and σ_{obs} are the corresponding standard deviations. This type of conditional bias occurs when the conditional expectation of the predictions depends on the observations, which is often a consequence of predictions with lower variance.

5. ANN predictions of growth and decay, verification, and predictability

a. Nowcasting of the mean growth and decay

Figure 4 serves as illustrative example and shows four prediction maps of the mean growth and decay. For demonstration purposes we used an ANN model with 5 input predictors [X , Y , U , V , HZT]. The trained ANN was asked to predict the GD fields for two different flow directions (SW, NW) and two different HZT (1500, 4000 m MSL) for a fixed flow speed of 30 km h^{-1} .

The prediction maps reproduce well known GD patterns in the Swiss Alps. As expected, precipitation growth is generally located on the northern slopes of the Alpine chain with NW flows, while with SW flows it is located on the southern side. A notable exception is the

region of growth upstream of the Berner Prealps with SW flows and high HZT (Fig. 4d). In contrast, the regions of decay are generally located in the inner Alpine chain and downstream with respect to the flow direction. All the main spatial patterns are in agreement with the radar-based climatology of Foresti et al. (2018).

The sensitivity of GD patterns on a given predictor can be studied by changing it in small steps and by fixing all the other predictors to a certain value (Andersen et al. 2017). This is an effective way to explore the radar-based climatology without defining arbitrary weather types. Animations of GD fields under different input conditions illustrate the full potential of the machine learning-based climatology; see some examples in the online supplemental material.

b. Verification of the growth and decay climatology

In this section, we analyze whether the ANN is able to reproduce the long-term GD climatology.

Following Foresti et al. (2018), we derived fields of average GD for different ranges of flow directions. The averages were computed for both the observed and the predicted GD on the whole 10-yr archive. The predictions were carried out based on the trained ANN of section 5a.

Figure 5 illustrates the working of the verification procedure for the class of SW flows ($225^\circ \pm 22.5^\circ$), independent on HZT. The top panel is the average GD field of ANN predictions, the center panel the average of GD observations and the third panel the average field of differences. It can be seen that the ANN climatology reproduces the observed climatology well, but has a slight tendency to underestimate (overestimate) the high (low) GD values.

Figure 6 shows the verification histograms of the mean observed GD against the mean predicted GD stratified by flow direction. The NW, SW, and W directions are well reproduced, with a RMSE below 0.3 dB and a PCORR above 0.95. However, the rarer flow conditions show a lower correspondence due to the small sample size (e.g., NE, E, and SE).

In summary, the ANN is able to learn and reproduce the GD climatology as a function of input predictors. However, the good correspondence of the average values does not imply a good correspondence of each instantaneous prediction of GD, which is much less predictable (see next section).

c. Analysis of the importance of external predictors

Figure 7 shows the verification of instantaneous GD predictions on the training, validation and test sets using different combinations of external predictors

FIG. 5. Example of climatological verification of ANN predictions for SW flows ($225^\circ \pm 22.5^\circ$). (top) Mean predicted GD (GD_{pred}); (middle) mean observed GD (GD_{obs}); and (bottom) mean GD differences ($GD_{diff} = GD_{pred} - GD_{obs}$). The ANN was trained with the predictors $[X, Y, U, V, HZT]$; N is the average number of samples per grid point.

$[X, Y, U, V, HZT, D_{sin}, D_{cos}]$. All the experiments were done using an ANN with 100 hidden neurons and a dataset size of 1 744 929 samples. The verification scores are computed by comparing each instantaneous GD prediction with the corresponding GD observation. It is important to note that the reported forecast performance is averaged over all spatial locations and times. Also, throughout the paper we will loosely use the terms

ANN vs observed climatology

FIG. 6. 2D histograms of long-term averages of observed and predicted GD. The example of Fig. 5 with SW flows is shown in the second row, third column. The RMSE, regression slope β , and PCORR are computed by comparing the $\overline{\text{GD}}_{\text{obs}}$ to the $\overline{\text{GD}}_{\text{pred}}$.

accuracy and *performance* to indicate forecasts with low RMSE and high PCORR, the latter being a measure of potential skill (Murphy 1995).

The first row in Fig. 7 is a two-input ANN model using only the geographical coordinates as predictors and has a RMSE of 2.8 dB and a PCORR of 0.21 on the test set. This experiment gives an estimation of the baseline performance when no additional information beyond the spatial location is available.

Rows 2–4 illustrate the results by adding the time of the day, the HZT, and the $[U, V]$ flow vectors to the

spatial coordinates (one at a time). Out of the three, the $[U, V]$ vectors have the greatest impact on the prediction performance, which is another evidence of the strong dependence of precipitation GD on flow direction and speed in the Alpine region (see Foresti et al. 2018). Out of the three, the time of the day has the least predictive power.

Rows 5–6 reveal that combining all external predictors, with and without time of the day, leads to negligible differences. The PCORR is around 0.29–0.30, and the RMSE is 2.72–2.73 dB. As expected, the reduction of RMSE is not as substantial as the increase of

FIG. 7. Analysis of the importance of external predictors to predict GD. Verification of ANN predictions on the training, validation, and test sets using the RMSE and the PCORR. On the y-axis there is the list of input predictors used, which are sorted from the top by increasing forecast quality.

PCORR, which can be attributed to the large contribution of the variance to the RMSE.

Given the small contribution of the time of day to prediction skill, in the following we will only work with the predictors $[X, Y, U, V, HZT]$.

Using the same set of predictors, we compared the predictive performance of the ANN, DT, and random forests. The skill being very similar, we decided to only include the results of ANN, as it also provides the most realistic growth and decay fields in terms of spatial continuity.

d. Application to nowcasting: Can machine learning improve beyond the persistence assumption?

Figure 8 employs 2D histograms to analyze the persistence of both the precipitation (MAP) and the GD using the whole 10-yr archive. Figures 8a and 8b show that the MAP is more persistent in the Lagrangian than in the Eulerian frame with a PCORR of 0.78 and 0.66, respectively. Figure 8c shows that the Eulerian persistence of GD only has a PCORR of 0.28, much lower than the one of MAP (see also Radhakrishna et al. 2012). The Eulerian persistence of GD reflects the stationary character of GD patterns over orography, which can be exploited to improve the prediction performance of the ANN.

Finally, in Fig. 8d we can observe a relationship between MAP_o and GD_d , which reveals an effect of *regression to the mean* (Barnett et al. 2005; Pitkänen et al. 2016). In essence, the larger a given MAP, the more likely

it is to decay and regress toward smaller values. The same applies to low MAP values but in the reverse sense.

Starting from these findings, we will answer the following questions:

- 1) Do the machine-learning-based predictions provide better performance than assuming persistence?
- 2) What is the impact of adding persistence information to the set of external predictors?
- 3) What is the impact of conditioning the predictions to the mean areal precipitation at the origin?

Figure 9 illustrates the results of the experiments designed to answer these three questions.

Row 1 shows the verification statistics by assuming Eulerian persistence of the GD values (same as Fig. 8c but for the training, validation and test sets). The PCORR of the different sets is in the range 0.28–0.30, while the RMSE is quite large and around 3.4–3.5 dB.

Row 2 depicts the base machine learning model using the set of 5 external predictors (same as row 5 in Fig. 7 but for a different training run). The PCORR of machine learning is only slightly larger than the one of Eulerian persistence, but the RMSE is substantially smaller (2.7 dB). This lower RMSE partly arises from the smoother machine learning predictions. Without considering PCORR, we would falsely conclude that machine learning provides much better accuracy than Eulerian persistence (question 1). It is worth pointing out that these

FIG. 8. 2D histograms to analyze the persistence of MAP and GD as well as their dependence. (a) Eulerian persistence of MAP, (b) Lagrangian persistence of MAP, (c) Eulerian persistence of GD, and (d) MAP_o vs GD. The regression line is obtained by a classical ordinary least squares fit with errors only in the y variable.

statements are only valid at the analyzed spatial (64 km) and temporal scales (1 h).

Row 3 shows an experiment using the spatial coordinates and the current GD as predictors, which reaches a PCORR of 0.34–0.35. Hence, learning the localized dependence structure of the GD based on the radar archive seems to be better than the persistence assumption.

Row 4 helps answer question 2 by using both the external predictors and the current GD to also exploit its persistence. Surprisingly, the PCORR is around 0.37–0.38, which is substantially higher than using either persistence (0.28–0.30) or the set of external predictors (0.30–0.31). Thus, we can enhance a persistence nowcast of GD by learning from the historical radar archive in combination with external predictors.

Row 5 answers question 3 by analyzing the impact of using MAP_o as an additional predictor. The increase in PCORR to 0.44–0.45 and decrease of RMSE benefits from the dependence of GD with

MAP_o (as shown in Fig. 8d), but the effect of regression to the mean questions whether the increased accuracy is real or merely a statistical artifact. This statistical property also has practical implications for operational nowcasting and warnings since using the MAP_o as predictor will have tendency to reduce the high MAP values and thus miss the extreme events. In such setting, it becomes essential to perform probabilistic predictions to allow the GD to increase with a certain probability, also when starting at high MAP values. This can be achieved by using prediction intervals as will be explained in section 6.

Finally, row 6 includes all the external and endogenous predictors, which brings the PCORR beyond 0.5 and the RMSE below 2.5 dB. This is a remarkable performance considering the fact that we are predicting the first time derivative of moving precipitation fields.

FIG. 9. Analysis of the impact of adding endogenous predictors to predict GD. MAP_o is given in dB units $[10 \log_{10}(MAP_o)]$.

e. Is it better to predict the growth and decay or to directly predict the precipitation intensity?

This section compares the following two settings:

- 1) Using ANN to predict the GD and derive the MAP_d (dB) as MAP_o (dB) + GD_{pred} (dB) [see (7)].
- 2) Using ANN to directly predict the MAP_d (dB) using the MAP_o (dB) and the set of external predictors.

Figure 10 illustrates the results of the two experiments. Row 1 shows the performance of Lagrangian persistence, which has a RMSE of 3 dB and a PCORR of 0.78 (same as Fig. 8b but for training, validation, and test).

Row 2 predicts MAP_d by adding the last observed GD to MAP_o . The PCORR is slightly reduced from 0.78 to 0.76 and the RMSE increases from 3 to 3.5 dB. The large variance of observed GD values could explain the increase of the RMSE.

Row 3 computes the MAP_d by adding to the MAP_o the GD predicted using the set of external predictors. With respect to Lagrangian persistence there is now a slight increase in prediction performance with a PCORR of 0.8 and a RMSE of 2.8 dB.

Row 4 is the same but additionally uses $GD_d(t)$ as predictor. The RMSE is reduced further to 2.75 dB and the PCORR rises to 0.81–0.82. This represents an $\approx 8\%$ decrease in RMSE and $\approx 5\%$ increase in PCORR with respect to Lagrangian persistence.

Figure 11 shows maps of the RMSE reduction by ANN for the NW and SW flows. As expected, the average 8% reduction of RMSE exhibits significant spatial variability depending on flow direction, but the ANN

never degrades the persistence-based nowcast. Over orography, the reduction is up to 15–20% in the regions of growth and 20–30% in the regions of decay.

Row 5 in Fig. 10 tests the effect of introducing MAP_o as a predictor. The RMSE is reduced further and the PCORR rises to 0.82–0.83. In this case, the skill with respect to Lagrangian persistence is $\approx 14\%$ for the RMSE and $\approx 6\%$ –7% for the PCORR.

Finally, row 6 shows the direct prediction of MAP_d using the same set of predictors of row 5. The performance is essentially indistinguishable from row 5. Thus, according to our experiments, there is no difference in performance between directly predicting the MAP_d or predicting the GD term and adding it to the MAP_o . However, there is a practical advantage in predicting GD as it simplifies the sensitivity analysis of input predictors (section 5a). This analysis would be more difficult by directly predicting MAP_d . In fact, it would require choosing an appropriate value for MAP_o to avoid confusing the sensitivity analysis with the effect of regression to the mean.

f. Verification scatterplots

Figure 12 shows the 2D verification histograms on the test set to better understand the effect of adding GD to the MAP and the regression to the mean.

Figures 12a and 12b show the verification for GD predictions without and with MAP_o as predictor. One can easily recognize that in both cases the range of GD predictions is much smaller than the one of the observations. This behavior is a natural consequence of

ANN verification of direct and indirect MAP_d prediction

FIG. 10. Verification results for direct and indirect prediction of MAP_d . To be consistent with the multiplicative formulation of GD, MAP_o and MAP_d are in dBR units.

the low predictability of GD and is observed as a conditional bias with respect to observations (regression slope β different from 1). The larger the departure of β is from 1 the stronger is the conditional bias. Such bias is unavoidable for any machine learning or other statistical model that tries to predict highly unpredictable atmospheric variables by minimizing the MSE (e.g., [Frei and Isotta 2019](#)). As already mentioned, a possible solution is to leave the deterministic world and perform probabilistic predictions.

[Figure 12b](#) also shows that adding MAP_o as predictor reduces the RMSE, increases the PCORR, and decreases the conditional bias. However, this conclusion is different when verifying the MAP_d prediction by adding

the GD to the MAP_o ([Fig. 12d](#)). In this case, the decrease of RMSE is not followed by a reduction of the conditional bias and the β deteriorates from 0.80 to 0.68. This can be noticed by the lower number of samples in both the lower-left and upper-right corners of the 2D histogram, where the range of MAP predictions is shrunk. This response is a direct consequence of the regression to the mean, which prevents high MAP values to grow further and generally increases the very low MAP values.

In conclusion, the effect of regression to the mean can reduce the RMSE, but can lead to a larger conditional bias. This statement is also valid for the direct prediction of MAP_d instead of GD (row 6 in [Fig. 10](#)), which yields the same $\beta = 0.68$ (not shown).

FIG. 11. Maps of RMSE reduction (%) by ANN with respect to Lagrangian persistence when predicting the MAP_d as $MAP_o + GD_{pred}$ using predictors $[X, Y, U, V, HZT, GD(t)]$ (row 4 in [Fig. 10](#)). (a) NW flows and (b) SW flows.

FIG. 12. Verification histograms on the test set for GD predictions: (a) without using MAP_o and (b) using MAP_o as predictor. Indirect prediction of MAP by adding the GD_{pred} : (a) without using MAP_o and (b) using MAP_o .

6. Probabilistic machine learning and quantification of prediction uncertainty

a. Prediction interval estimation in machine learning

The topic of uncertainty quantification in machine learning has received increasing attention in recent years (see e.g., Ghahramani 2015). The uncertainty can be estimated by computing prediction and confidence intervals (Heskes 1997):

- The *prediction interval* (PI) estimates the range in which a new observation will fall with a certain probability. It measures the uncertainty of predictions.
- The *confidence interval* (CI) estimates the range in which a model parameter will fall with a certain probability. It measures the uncertainty of parameters.

The CI is contained in the PI and is usually much smaller. For instance, the spread of an NWP ensemble can be used as an estimate of PI.

There are several ways to estimate the PI depending on the chosen ML algorithm and adopted philosophy

(i.e., Bayesian or frequentist) (see e.g., Heskes 1997; Meinshausen 2006; Khosravi et al. 2011; Ghahramani 2015). ANN-based approaches typically derive the PI by training an ensemble of ANNs on bootstrap replicates of the training set and/or by fitting an ANN model to the squared residuals of the validation set (Heskes 1997; Khosravi et al. 2011). The bootstrapping approach is only feasible with small datasets (see e.g., Khosravi et al. 2011), while fitting the squared residuals implies assuming a Gaussian distribution of the errors. To relax this assumption, one can approximate the distribution by estimating a finite set of quantiles (e.g., Cade and Noon 2003), for example using a dedicated ANN for each quantile (e.g., Cannon 2011).

Decision trees can easily be extended to perform quantile regression (Meinshausen 2006). A naïve quantile decision tree can be devised by computing a set of empirical quantiles from the collection of target values at each leaf. Random forests can be used to further stabilize the quantile estimations. The drawback of

tree-based methods is that they provide a step-wise estimation of the regression function.

b. A new method for prediction interval estimation

QUANTILE ANN BASED ON DECISION TREES

To benefit from the computational speed of decision trees and the smoothness of ANN predictions, we propose a combined approach for prediction interval estimation: *quantile neural network based on decision trees* (QANN). The idea is as follows (see Fig. 13):

- 1) Train a decision tree (or random forest) to predict the target variable. Optimize the model hyperparameters by using the validation dataset.
- 2) Loop over each leaf of the tree and compute a set of quantiles (e.g., 1%, 5%, 25%, 50%, 75%, 95%, 99%), from the target values of the training and validation sets.
- 3) Train a multioutput MLP using as input the same predictors as point 1 and as outputs the quantiles of the training set derived at point 2. Use the validation error for early stopping.

For quantile regression we use the python library “scikit-garden” (skgarden), which computes weighted percentiles (see <https://scikit-garden.github.io>). We also extended the library to return the first four moments of the distribution (i.e., the mean, variance, skewness, and kurtosis). This setting can be used, for example, to decompose the total model errors (MSE) into bias and variance components.

Instead of training a single decision tree at point 1., one could alternatively train a random forest and compute average quantiles from the trees. However, since our dataset is quite large, using a single decision tree represents a sufficiently good first approximation. Finally, with small datasets, it may be better to interpolate the quantiles of the validation instead of the training set to avoid overfitting (see e.g., Heskes 1997).

A possible improvement of QANN could be to extend the error function of the ANN to ensure that the quantiles cannot cross each other. However, our experiments show that in our case, the ANN interpolation fully preserves the ranking of the decision tree quantiles (see section 7b), which are monotonously increasing by definition. Though, we expect that crossings would be more likely to occur only if we asked the QANN to predict very close quantiles.

QANN follows the philosophy of Solomatine and Shrestha (2009), who employ machine learning models to learn the dependence between the input set of predictors and the historical model errors. In this study, we use QANN to highlight that the estimation of nowcast

FIG. 13. Example of multioutput MLP used for prediction interval estimation (QANN). The target quantiles are previously computed based on decision trees and are passed as target values to the MLP.

uncertainty is as important as the estimation of the conditional mean. Further studies could focus on finding improved QANN settings and comparing them with more mature algorithms, such as quantile random forests (Meinshausen 2006) or quantile neural networks (Cannon 2011).

c. Testing QANN on a numerical example

The new algorithm is tested with simulated data (see Fig. 14a). The true output values are generated by the function $f(x) = x \sin(x)$, which is perturbed with an increasing heteroskedastic noise term. We trained two models: the naïve quantile decision trees and the QANN. Both algorithms are trained to predict the quantiles 1%, 5%, 25%, 50%, 75%, 95%, and 99% using 350 points for training, 150 for validation, and 1000 for testing. In this case, the MLP has 1 input neuron, 50 hidden neurons, and 7 output neurons.

Figure 14a shows the prediction results together with the 90% PI (i.e., the range 5%–95%). Quantile DT are able to capture the heteroskedastic behavior of the noise term (left panel), but the prediction function is stepwise (discontinuous). In fact, the conditional statistical moments and quantiles are assumed to be constant within each leaf. As the quantiles are computed empirically from the target values, they also have tendency to overfit the data [see e.g., the point at $x, f(x) = (-1.7, -12)$]. QANN also captures the heteroskedastic behavior of the noise term (right panel), but produces a smooth prediction interval. Additionally, it is more robust to data outliers.

One limitation of the QANN is the overestimation of the PI width in regions with strong gradients of

FIG. 14. Numerical experiment to test the QANN method. (a) Prediction results for the 90% PI interval using (left) the quantile DT and (right) the QANN; the RMSE and PCORR measure the correspondence between the observed values and the conditional median (red line). (b) Corresponding verification of the predicted quantiles using a reliability diagram, where the observed frequency below a certain quantile is plotted against the predicted frequency; the RMSE measures the average error between the target quantiles of the decision tree and the ones predicted by the QANN. The selected hyperparameters are also shown.

the target variable [e.g., on the right side of the function ($x > 1.0$)]. This is due to the DT assuming a constant mean value within the leaf. One solution could be to perform a stepwise linear regression of the target values at the cost of additional computational time.

Figure 14b illustrates the verification results for the two approaches with a reliability plot on the training, validation, and test sets. The correspondence of observed and predicted frequencies of QANN is better than the quantile DT. The proportions of observations falling in the 50%, 90%, and 98% PI are also better

reproduced by QANN, which are 46.4%, 89.1%, and 94.4%, respectively.

7. Using QANN to estimate the uncertainty of growth and decay predictions

a. Prediction interval estimation of growth and decay

Figure 15 uses the same predictors as Fig. 4, but instead of predicting the mean GD it uses QANN to predict the 90% prediction interval.

The QANN is able to capture the larger GD uncertainty associated with high HZT conditions. In fact, with

FIG. 15. QANN predictions of the growth and decay 90% prediction interval with different flow directions and freezing level heights for a fixed flow speed of 30 km h^{-1} . NW flows with (a) HZT at 1500 m and (b) HZT at 4000 m. SW flows with (c) HZT at 1500 m and (d) HZT at 4000 m.

high HZT the 90% PI interval is in the range 8–15 dB, while with low HZT it is in the range 7–10 dB. It is also interesting to note that the PI is larger on the western side of the domain with high HZT, which represents the lower predictability of growth and decay over the flat areas of France compared to the Alpine region. Also, the spatial patterns of PI display a lower spatial variability and dependence with flow direction compared to the conditional mean GD predictions of the ANN (Fig. 4).

Finally, it is important to mention that even a field with a mean GD ≈ 0 everywhere (e.g., over flat continental regions) can still exhibit variability, and thus predictability, of the prediction interval. Hence, additional information about predictability is gained, which could not have been found by only modeling the conditional mean (e.g., Cade and Noon 2003).

b. Verification of growth and decay quantiles

Figure 16 illustrates the verification of the quantiles predicted by the QANN of previous section.

Figure 16a shows an almost perfect correspondence between the observed and predicted quantiles, which is even better than the one of the numerical example (section 6c), probably a consequence of the larger sample size. This demonstrates that the predicted quantiles are well calibrated (unbiased). Their discrimination ability, however, is limited by the prediction performance of the decision tree. Finally, no crossing quantiles were found in the training, validation, and test sets, which confirms that the ANN fully preserves the ranking of decision tree quantiles.

In Fig. 16b we rank a random subset of observations of the test set by increasing PI width. The lowest values of the 98% PI are around 10 dB while the highest around 20 dB, which reflects again the low predictability of precipitation GD and the importance of estimating the forecast uncertainty. On average, we should expect 2% of the observations to fall outside the 98% PI, respectively, 10% outside the 90% PI. Indeed, there are 9 out of 500 points falling outside the 98% PI, which corresponds to 1.8%.

FIG. 16. Verification of the predicted quantiles by QANN on the growth and decay dataset. (a) Reliability plot showing the observed vs predicted quantiles and the percentage of values falling within a given PI. (b) Plot displaying a subset of observed GD values on the test set ranked by increasing 98% PI width. The PI values are centered to remove the variations of the median and improve the clarity of the plot as in Meinshausen (2006).

8. Discussion

a. Relationship between the machine learning and the analog approach

In this study, we used machine learning to extract the predictable precipitation patterns from a historical radar data archive. A closely related approach is the concept of analogs, which assumes that the current weather situation will evolve similarly as it did in analog situations in the past (Lorenz 1969; Toth 1991b). Analog-based radar nowcasting studies can be found in Panziera et al. (2011); Foresti et al. (2015); Atencia and Zawadzki (2015). Both machine learning and analog approaches start from the same dataset and suffer from the same limitations due to its finite size (see e.g., Toth 1991a; Van Den Dool 1994).

One solution to increase the probability of finding similar atmospheric states is to localize the search of analogs to smaller domains, as done in the *local analog approach* (see e.g., Hamill and Whitaker 2006; Li and Ding 2011). Using geographical coordinates as input predictors for an ANN represents an interesting solution to retrieve local analogs while preserving the continuity of the field. Moreover, using the current GD and MAP as predictors can help imposing spatial coherence to the predicted GD fields.

Nevertheless, it is important to understand that the fields predicted by machine learning methods are not realizations of the future state of the atmosphere, but merely their statistical moments (mean, variance, quantiles, etc.). For instance, predicting the conditional mean with machine learning is similar to computing the ensemble mean of a set of analogs. As such, machine learning simply performs an interpolation of analog states so as to minimize a chosen error function.

In addition to prediction interval estimation, stochastic simulation could be used to generate a set of realistic ensemble members that honor the statistical moments and reproduce the correct space-time correlations of the probability density function (e.g., Bowler et al. 2006; Germann et al. 2009; Nerini et al. 2017; Frei and Isotta 2019). An interesting way to perform both tasks at the same time (i.e., error minimization and ensemble generation) is to use generative adversarial neural networks, as shown by Gagne et al. (2018).

b. Postprocessing of nowcasts and NWP forecasts

The GD term represents the forecast error by Lagrangian persistence, where the current precipitation value (MAP_o) is the persistence-based nowcast and the next precipitation value (MAP_d) is the verifying observation. The methodology of this paper could also be applied for NWP postprocessing (e.g., by using QANN to estimate the NWP model errors with respect to the measurements at weather stations; Taillardat et al. 2016; Rasp and Lerch 2018). Using geographical coordinates as predictors would give a natural way to interpolate the model errors between the weather stations (see e.g., Weingart 2018).

The postprocessing of precipitation nowcasts opens several interesting possibilities. We currently see three main ways to estimate the growth and decay:

- Estimating GD using statistical learning or analog approaches from the radar data archives.
- Estimating GD from the observed sequence of radar rainfall fields (e.g., the last 2–3 h) and assume Eulerian and/or Lagrangian persistence (e.g., Sideris et al. 2018; Radhakrishna et al. 2012).
- Estimating GD from the forecasted sequence of NWP rainfall fields (e.g., as done by Sideris et al. 2018).

An interesting question is to understand how to blend the growth and decay of the real-time radar and NWP fields with the one derived from the archive. Machine learning could provide a flexible framework to optimally integrate these different data sources.

9. Conclusions

We presented a machine learning framework for nowcasting precipitation growth and decay in the Swiss Alpine region based on a 10-yr archive of composite radar images. The trained artificial neural networks were able to automatically learn and reproduce the climatological growth and decay patterns, in agreement with the findings of Foresti et al. (2018).

Forecast verification revealed the most relevant predictors, which are in order of importance: the geographical location, the flow direction and speed, and the freezing level height. The ANN predictions provided similar accuracy as assuming persistence of growth and decay, but when combined with the latter the performance improved substantially. The decrease of RMSE compared with persistence is up to 20%–30% over orography.

Deterministic machine learning predictions are designed to minimize prediction errors, which, however, lead to smooth forecast fields characterized by strong conditional biases. This complicates the comparison of machine learning predictions with the persistence baseline, which, by definition, preserves the variance of observations. To overcome these limitations, we introduced a probabilistic machine learning framework for precipitation nowcasting and presented a novel method to estimate the prediction uncertainty based on a combination of decision trees and ANNs (i.e., QANN). Such uncertainty estimates could be used in combination with stochastic simulation (e.g., Nerini et al. 2017; Frei and Isotta 2019) to generate a realistic ensemble of precipitation fields. Future advances in machine learning should consider extending also the deep convolutional neural networks (e.g., Shi et al. 2015) to estimate the prediction uncertainty.

The analyses and conclusions of the paper are relative to a spatial scale of 64 km and a lead time of 1 h. Thus, an interesting extension could be to study the scale and lead-time dependence of the predictive performance. Another open question concerns the residual radar measurement uncertainty, which locally affects the growth and decay values (see a discussion in Foresti and Seed 2015; Foresti et al. 2018).

Additional radar, satellite, and NWP predictors could also be included to further enhance the prediction performance (e.g., Mecikalski et al. 2015; Han et al. 2017; Zeder et al. 2018). However, as the atmosphere is a chaotic system characterized by intrinsic predictability

limits, we do not expect large improvements (i.e., it will remain necessary to estimate the prediction uncertainty, e.g., by using prediction intervals).

The presented machine learning framework could readily be applied to derive a thunderstorm climatology using the large archives of convective cell tracks (e.g., Goudenhoofdt and Delobbe 2013; Meyer et al. 2013; Wapler and James 2015; Nisi et al. 2018). In fact, these datasets contain similar predictors, such as the spatial location of the cell $[X, Y]$, the tracked motion vectors $[U, V]$, and, potentially, NWP, satellite, and lightning variables describing the environmental conditions and life cycle of the storm. This analysis would not only be interesting from a climatological perspective, but could also form a basis to incorporate information about the evolution of individual convective cells into field-based nowcasting systems (Sideris et al. 2018).

Acknowledgments. This study was supported by the Swiss National Science Foundation Ambizione project “Precipitation attractor from radar and satellite data archives and implications for seamless very short-term forecasting” (PZ00P2 161316). We thank Ulrich Hamann, Alan Seed, Luca Panziera, Simona Trefalt, Marco Gabella and Floor van den Heuvel for the useful discussions and feedback on the manuscript. Bertrand Calpini is thanked for his support to the project. We are also grateful to Christoph Frei for the discussion on error minimization and conditional bias.

REFERENCES

- Andersen, H., J. Cermak, J. Fuchs, R. Knutti, and U. Lohmann, 2017: Understanding the drivers of marine liquid-water cloud occurrence and properties with global observations using neural networks. *Atmos. Chem. Phys.*, **17**, 9535–9546, <https://doi.org/10.5194/acp-17-9535-2017>.
- Atencia, A., and I. Zawadzki, 2014: A comparison of two techniques for generating nowcasting ensembles. Part I: Lagrangian ensemble technique. *Mon. Wea. Rev.*, **142**, 4036–4052, <https://doi.org/10.1175/MWR-D-13-00117.1>.
- , and —, 2015: A comparison of two techniques for generating nowcasting ensembles. Part II: Analogs selection and comparison of techniques. *Mon. Wea. Rev.*, **143**, 2890–2908, <https://doi.org/10.1175/MWR-D-14-00342.1>.
- , —, and M. Berenguer, 2017: Scale characterization and correction of diurnal cycle errors in MAPLE. *J. Appl. Meteor. Climatol.*, **56**, 2561–2575, <https://doi.org/10.1175/JAMC-D-16-0344.1>.
- Baldauf, M., A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer, and T. Reinhardt, 2011: Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Mon. Wea. Rev.*, **139**, 3887–3905, <https://doi.org/10.1175/MWR-D-10-05013.1>.
- Barnett, A., J. van der Pols, and A. Dobson, 2005: Regression to the mean: What it is and how to deal with it. *Int. J. Epidemiol.*, **34**, 215–220, <https://doi.org/10.1093/ije/dyh299>.

- Berenguer, M., D. Sempere-Torres, and G. G. Pegram, 2011: SBMcast: An ensemble nowcasting technique to assess the uncertainty in rainfall forecasts by Lagrangian extrapolation. *J. Hydrol.*, **404**, 226–240, <https://doi.org/10.1016/j.jhydrol.2011.04.033>.
- Besic, N., J. Figueras i Ventura, J. Grazioli, M. Gabella, U. Germann, and A. Berne, 2016: Hydrometeor classification through statistical clustering of polarimetric radar measurements: A semi-supervised approach. *Atmos. Meas. Tech.*, **9**, 4425–4445, <https://doi.org/10.5194/amt-9-4425-2016>.
- Beusch, L., L. Foresti, M. Gabella, and U. Hamann, 2018: Satellite-based rainfall retrieval: From generalized linear models to artificial neural networks. *Remote Sens.*, **10**, 939, <https://doi.org/10.3390/rs10060939>.
- Bowler, N. E., C. E. Pierce, and A. Seed, 2006: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quart. J. Roy. Meteor. Soc.*, **132**, 2127–2155, <https://doi.org/10.1256/qj.04.100>.
- Breiman, L., 2001a: Random forests. *Mach. Learn.*, **45**, 5–32, <https://doi.org/10.1023/A:1010933404324>.
- , 2001b: Statistical modeling: The two cultures. *Stat. Sci.*, **16**, 199–231, <https://doi.org/10.1214/ss/1009213726>.
- , J. H. Friedman, R. A. Olshen, and C. J. Stone, 1984: *Classification and Regression Trees*. Chapman and Hall/CRC, 368 pp.
- Cade, B., and B. Noon, 2003: A gentle introduction to quantile regression for ecologists. *Front. Ecol. Environ.*, **1**, 412–420, [https://doi.org/10.1890/1540-9295\(2003\)001\[0412:AGITQR\]2.0.CO;2](https://doi.org/10.1890/1540-9295(2003)001[0412:AGITQR]2.0.CO;2).
- Cannon, A., 2011: Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Comput. Geosci.*, **37**, 1277–1284, <https://doi.org/10.1016/j.cageo.2010.07.005>.
- Carbone, R., J. Tuttle, D. Ahijevych, and S. Trier, 2002: Inferences of predictability associated with warm season precipitation episodes. *J. Atmos. Sci.*, **59**, 2033–2056, [https://doi.org/10.1175/1520-0469\(2002\)059<2033:IOPAWW>2.0.CO;2](https://doi.org/10.1175/1520-0469(2002)059<2033:IOPAWW>2.0.CO;2).
- Fabry, F., and A. Seed, 2009: Quantifying and predicting the accuracy of radar-based quantitative precipitation forecasts. *Adv. Water Resour.*, **32**, 1043–1049, <https://doi.org/10.1016/j.advwatres.2008.10.001>.
- , V. Meunier, B. Treserras, A. Cournoyer, and B. Nelson, 2017: On the climatological use of radar data mosaics: Possibilities and challenges. *Bull. Amer. Meteor. Soc.*, **98**, 2135–2148, <https://doi.org/10.1175/BAMS-D-15-00256.1>.
- Foresti, L., and A. Seed, 2015: On the spatial distribution of rainfall nowcasting errors due to orographic forcing. *Meteor. Appl.*, **22**, 60–74, <https://doi.org/10.1002/met.1440>.
- , M. Kanevski, and A. Pozdnoukhov, 2012: Kernel-based mapping of orographic rainfall enhancement in the Swiss Alps as detected by weather radar. *IEEE Trans. Geosci. Remote Sens.*, **50**, 2954–2967, <https://doi.org/10.1109/TGRS.2011.2179550>.
- , L. Panziera, P. V. Mandapaka, U. Germann, and A. Seed, 2015: Retrieval of analogue radar images for ensemble nowcasting of orographic rainfall. *Meteor. Appl.*, **22**, 141–155, <https://doi.org/10.1002/met.1416>.
- , I. Sideris, L. Panziera, D. Nerini, and U. Germann, 2018: A 10-year radar-based analysis of orographic precipitation growth and decay patterns over the Swiss Alpine region. *Quart. J. Roy. Meteor. Soc.*, **144**, 2277–2301, <https://doi.org/10.1002/qj.3364>.
- Frei, C., and F. Isotta, 2019: Ensemble spatial precipitation analysis from rain gauge data - Methodology and application in the European Alps. *J. Geophys. Res. Atmos.*, **124**, 5757–5778, <https://doi.org/10.1029/2018JD030004>.
- French, M., W. Krajewski, and R. Cuykendall, 1992: Rainfall forecasting in space and time using a neural network. *J. Hydrol.*, **137**, 1–31, [https://doi.org/10.1016/0022-1694\(92\)90046-X](https://doi.org/10.1016/0022-1694(92)90046-X).
- Freund, Y., and R. Schapire, 1997: A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139, <https://doi.org/10.1006/jcss.1997.1504>.
- Fukushima, K., and S. Miyake, 1982: Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognit.*, **15**, 455–469, [https://doi.org/10.1016/0031-3203\(82\)90024-3](https://doi.org/10.1016/0031-3203(82)90024-3).
- Gagne, D., A. McGovern, S. Haupt, R. Sobash, J. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, <https://doi.org/10.1175/WAF-D-17-0010.1>.
- , S. Haupt, D. Nychka, H. Christensen, A. Subramanian, and A. Monahan, 2018: Generation of spatial weather fields with generative adversarial networks. *Fourth Conf. on Stochastic Weather Generators (SWGEn 2018)*, Boulder, CO, University Corporation for Atmospheric Research, <http://opensky.ucar.edu/islandora/object/conference:3343>.
- Germann, U., and I. Zawadzki, 2002: Scale-dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology. *Mon. Wea. Rev.*, **130**, 2859–2873, [https://doi.org/10.1175/1520-0493\(2002\)130<2859:SDOTPO>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<2859:SDOTPO>2.0.CO;2).
- , —, and B. Turner, 2006: Predictability of precipitation from continental radar images. Part IV: Limits to prediction. *J. Atmos. Sci.*, **63**, 2092–2108, <https://doi.org/10.1175/JAS3735.1>.
- , M. Berenguer, D. Sempere-Torres, and M. Zappa, 2009: REAL-Ensemble radar precipitation estimation for hydrology in a mountainous region. *Quart. J. Roy. Meteor. Soc.*, **135**, 445–456, <https://doi.org/10.1002/qj.375>.
- , D. Nerini, I. Sideris, L. Foresti, A. Hering, and B. Calpini, 2017: *Real-time radar - A new Alpine radar network*. Meteorological Technology Int., 4 pp., <https://www.meteosuisse.admin.ch/content/dam/meteoswiss/en/Mess-Prognosesysteme/Atmosphaere/doc/MTI-April2017-Rad4Alp.pdf>.
- Ghahramani, Z., 2015: Probabilistic machine learning and artificial intelligence. *Nature*, **521**, 452–459, <https://doi.org/10.1038/nature14541>.
- Glorot, X., A. Bordes, and Y. Bengio, 2011: Deep sparse rectifier neural networks. *Proc. Machine Learning Res.*, **15**, 315–323.
- Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. Adaptive Computation and Machine Learning Series, F. Bach, Ed., MIT Press, 800 pp., <http://www.deeplearningbook.org>.
- Goudenhoofd, E., and L. Delobbe, 2013: Statistical characteristics of convective storms in Belgium derived from volumetric weather radar observations. *J. Appl. Meteor. Climatol.*, **52**, 918–934, <https://doi.org/10.1175/JAMC-D-12-079.1>.
- Grecu, M., and W. Krajewski, 2000: A large-sample investigation of statistical procedures for radar-based short-term quantitative precipitation forecasting. *J. Hydrol.*, **239**, 69–84, [https://doi.org/10.1016/S0022-1694\(00\)00360-7](https://doi.org/10.1016/S0022-1694(00)00360-7).
- Hall, T., H. Brooks, and C. Doswell III, 1999: Precipitation forecasting using a neural network. *Wea. Forecasting*, **14**, 338–345, [https://doi.org/10.1175/1520-0434\(1999\)014<0338:PFUANN>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0338:PFUANN>2.0.CO;2).
- Hamill, T., and J. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, [10.1175/MWR3237.1](https://doi.org/10.1175/MWR3237.1).

- Han, L., J. Sun, W. Zhang, Y. Xiu, H. Feng, and Y. Lin, 2017: A machine learning nowcasting method based on real-time re-analysis data. *J. Geophys. Res. Atmos.*, **122**, 4038–4051, <https://doi.org/10.1002/2016JD025783>.
- Haupt, S., A. Pasini, and C. Marzban, Eds., 2009: *Artificial Intelligence Methods in the Environmental Sciences*. Springer, 424 pp., <https://doi.org/10.1007/978-1-4020-9119-3>.
- Haykin, S., 1998: *Neural Networks: A Comprehensive Foundation*. 2nd ed. Prentice-Hall, 842 pp.
- Heskes, T., 1997: Practical confidence and prediction intervals. *Adv. Neural Info. Process. Syst.*, **9**, 176–182.
- Hinton, G. E., S. Osindero, and Y. Teh, 2006: A fast learning algorithm for deep belief nets. *Neural Comput.*, **18**, 1527–1554, <https://doi.org/10.1162/neco.2006.18.7.1527>.
- Kanevski, M., V. Timonin, and A. Pozdnoukhov, 2009: *Machine Learning for Spatial Environmental Data: Theory, Applications, and Software*. EPFL Press, 400 pp.
- Khosravi, A., S. Nahavandi, D. Creighton, and A. Atiya, 2011: Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Trans. Neural Network*, **22**, 1341–1356, <https://doi.org/10.1109/TNN.2011.2162110>.
- Kingma, D., and J. Ba, 2015: Adam: A method for stochastic optimization. *Third Int. Conf. on Learning Representations (ICLR 2015)*, San Diego, CA, ICLR, <http://arxiv.org/abs/1412.6980>.
- Kretzschmar, R., P. Eckert, D. Cattani, and F. Eggimann, 2004: Neural network classifiers for local wind prediction. *J. Appl. Meteor.*, **43**, 727–738, <https://doi.org/10.1175/2057.1>.
- Kuligowski, R., and A. Barros, 1998: Experiments in short-term precipitation forecasting using artificial neural networks. *Mon. Wea. Rev.*, **126**, 470–482, [https://doi.org/10.1175/1520-0493\(1998\)126<0470:EISTPF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<0470:EISTPF>2.0.CO;2).
- Li, J., and R. Ding, 2011: Temporal-spatial distribution of atmospheric predictability limit by local dynamical analogs. *Mon. Wea. Rev.*, **139**, 3265–3283, <https://doi.org/10.1175/MWR-D-10-05020.1>.
- Lorenz, E. N., 1956: *Empirical orthogonal functions and statistical weather prediction*. Department of Meteorology, Massachusetts Institute of Technology, 52 pp.
- , 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, [https://doi.org/10.1175/1520-0469\(1969\)26<636:APARBN>2.0.CO;2](https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2).
- , 1996: Predictability—A problem partly solved. *Proc. Seminar on Predictability*, Vol. 1, Reading, Berkshire, United Kingdom, ECMWF, 18 pp., <https://www.ecmwf.int/en/elibrary/10829-predictability-problem-partly-solved>.
- Malone, T., 1955: Applications of statistical methods in weather prediction. *Proc. Natl. Acad. Sci. USA*, **41**, 806–815, <https://doi.org/10.1073/pnas.41.11.806>.
- Mandapaka, P., U. Germann, and L. Panziera, 2013: Diurnal cycle of precipitation over complex alpine orography: Inferences from high-resolution radar observations. *Quart. J. Roy. Meteor. Soc.*, **139**, 1025–1046, <https://doi.org/10.1002/qj.2013>.
- Manzato, A., 2005: The use of sounding-derived indices for a neural network short-term thunderstorm forecast. *Wea. Forecasting*, **20**, 896–917, <https://doi.org/10.1175/WAF898.1>.
- Marzban, C., and A. Witt, 2001: A Bayesian neural network for severe-hail size prediction. *Wea. Forecasting*, **16**, 600–610, [https://doi.org/10.1175/1520-0434\(2001\)016<0600:ABNNFS>2.0.CO;2](https://doi.org/10.1175/1520-0434(2001)016<0600:ABNNFS>2.0.CO;2).
- McCann, D., 1992: A neural network short-term forecast of significant thunderstorms. *Wea. Forecasting*, **7**, 525–534, [https://doi.org/10.1175/1520-0434\(1992\)007<0525:ANNSTF>2.0.CO;2](https://doi.org/10.1175/1520-0434(1992)007<0525:ANNSTF>2.0.CO;2).
- McGovern, A., D. Elmore, K. L. Gagne, S. Haupt, C. Karstens, R. Lagerquist, T. Smith, and J. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, <https://doi.org/10.1175/BAMS-D-16-0123.1>.
- Mecikalski, J., J. Williams, C. Jewett, D. Ahijevych, A. LeRoy, and J. Walker, 2015: Probabilistic 0–1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *J. Appl. Meteor. Climatol.*, **54**, 1039–1059, <https://doi.org/10.1175/JAMC-D-14-0129.1>.
- Meinshausen, N., 2006: Quantile regression forests. *J. Mach. Learn. Res.*, **7**, 983–999.
- Meyer, V., H. Höller, and H. Betz, 2013: Automated thunderstorm tracking: Utilization of three-dimensional lightning and radar data. *Atmos. Chem. Phys.*, **13**, 5137–5150, <https://doi.org/10.5194/acp-13-5137-2013>.
- Murphy, A. H., 1995: The coefficients of correlation and determination as measures of performance in forecast verification. *Wea. Forecasting*, **10**, 681–688, [https://doi.org/10.1175/1520-0434\(1995\)010<0681:TCOCAD>2.0.CO;2](https://doi.org/10.1175/1520-0434(1995)010<0681:TCOCAD>2.0.CO;2).
- Nerini, D., N. Besic, I. Sideris, U. Germann, and L. Foresti, 2017: A non-stationary stochastic ensemble generator for radar rainfall fields based on the short-space Fourier transform. *Hydrol. Earth Syst. Sci.*, **21**, 2777–2797, <https://doi.org/10.5194/hess-21-2777-2017>.
- Nisi, L., A. Hering, U. Germann, and O. Martius, 2018: A 15-year hail streak climatology for the Alpine region. *Quart. J. Roy. Meteor. Soc.*, **144**, 1429–1449, <https://doi.org/10.1002/qj.3286>.
- Panziera, L., U. Germann, M. Gabella, and P. V. Mandapaka, 2011: NORA—Nowcasting of orographic rainfall by means of analogues. *Quart. J. Roy. Meteor. Soc.*, **137**, 2106–2123, <https://doi.org/10.1002/qj.878>.
- Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
- Peleg, N., F. Marra, S. Fatichi, P. Molnar, E. Morin, A. Sharma, and P. Burlando, 2018: Intensification of convective rain cells at warmer temperatures observed from high-resolution weather radar data. *J. Hydrometeorol.*, **19**, 715–726, <https://doi.org/10.1175/JHM-D-17-0158.1>.
- Pitkänen, M., S. Mikkonen, K. Lehtinen, A. Lipponen, and A. Arola, 2016: Artificial bias typically neglected in comparisons of uncertain atmospheric data. *Geophys. Res. Lett.*, **43**, 10 003–10 011, <https://doi.org/10.1002/2016GL070852>.
- Potts, J., 2012: Basic concepts. *Forecast Verification: A Practitioner's Guide in Atmospheric Sciences*, I. T. Jolliffe and D. B. Stephenson, Eds., Wiley-Blackwell, 11–29.
- Radhakrishna, B., I. Zawadzki, and F. Fabry, 2012: Predictability of precipitation from continental radar images. Part V: Growth and decay. *J. Atmos. Sci.*, **69**, 3336–3349, <https://doi.org/10.1175/JAS-D-12-029.1>.
- Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, <https://doi.org/10.1175/MWR-D-18-0187.1>.
- Schmidhuber, J., 2015: Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117, <https://doi.org/10.1016/j.neunet.2014.09.003>.
- Seed, A., 2003: A dynamic and spatial scaling approach to advection forecasting. *J. Appl. Meteor.*, **42**, 381–388, [https://doi.org/10.1175/1520-0450\(2003\)042<0381:ADASSA>2.0.CO;2](https://doi.org/10.1175/1520-0450(2003)042<0381:ADASSA>2.0.CO;2).
- , C. E. Pierce, and K. Norman, 2013: Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme. *Water Resour. Res.*, **49**, 6624–6641, <https://doi.org/10.1002/wrcr.20536>.

- Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, 2015: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Info. Process. Syst.*, **28**, 802–810, <https://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowcasting.pdf>.
- Sideris, I., L. Foresti, D. Nerini, and U. Germann, 2018: Now-Precip: An algorithm for localized probabilistic precipitation nowcasting in the complex terrain of Switzerland. *Proc. 10th European Conf. on Radar in Meteorology and Hydrology (ERAD)*, Ede-Wageningen, the Netherlands, ERAD, Abstract number 192, 9 pp., http://projects.knmi.nl/erad2018/ERAD2018_extended_abstract_192.pdf.
- Solomatine, D., and L. Shrestha, 2009: A novel method to estimate model uncertainty using machine learning techniques. *Water Resour. Res.*, **45**, 1–6, <https://doi.org/10.1029/2008WR006839>.
- Sprenger, S., S. Schemm, R. Oechslin, and J. Jenkner, 2017: Nowcasting Foehn wind events using the AdaBoost machine learning algorithm. *Wea. Forecasting*, **32**, 1079–1099, <https://doi.org/10.1175/WAF-D-16-0208.1>.
- Surcel, M., I. Zawadzki, and M. K. Yau, 2015: A study on the scale dependence of the predictability of precipitation patterns. *J. Atmos. Sci.*, **72**, 216–235, <https://doi.org/10.1175/JAS-D-14-0071.1>.
- Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, <https://doi.org/10.1175/MWR-D-15-0260.1>.
- Toth, Z., 1991a: Circulation patterns in phase space: A multi-normal distribution? *Mon. Wea. Rev.*, **119**, 1501–1511, [https://doi.org/10.1175/1520-0493\(1991\)119<1501:CPIPSA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1991)119<1501:CPIPSA>2.0.CO;2).
- , 1991b: Estimation of atmospheric predictability by circulation analogues. *Mon. Wea. Rev.*, **119**, 65–72, [https://doi.org/10.1175/1520-0493\(1991\)119<0065:EOAPBC>2.0.CO;2](https://doi.org/10.1175/1520-0493(1991)119<0065:EOAPBC>2.0.CO;2).
- Tsonis, A., and G. Austin, 1981: An evaluation of extrapolation techniques for the short-term prediction of rain amounts. *Atmos.–Ocean*, **19**, 54–65, <https://doi.org/10.1080/07055900.1981.9649100>.
- Turner, B., I. Zawadzki, and U. Germann, 2004: Predictability of precipitation from continental radar images. Part III: Operational nowcasting implementation (MAPLE). *J. Appl. Meteor.*, **43**, 231–248, [https://doi.org/10.1175/1520-0450\(2004\)043<0231:POPFGR>2.0.CO;2](https://doi.org/10.1175/1520-0450(2004)043<0231:POPFGR>2.0.CO;2).
- Ukkonen, P., A. Manzato, and A. Määkelä, 2017: Evaluation of thunderstorm predictors for Finland using reanalyses and neural networks. *J. Appl. Meteor. Climatol.*, **56**, 2335–2352, <https://doi.org/10.1175/JAMC-D-16-0361.1>.
- Van Den Dool, H. M., 1994: Searching for analogs: How long must we wait? *Tellus*, **46A**, 314–324, <https://doi.org/10.3402/tellusa.v46i3.15481>.
- Villarini, G., and W. F. Krajewski, 2010: Review of the different sources of uncertainty in single polarization radar-based estimates of rainfall. *Surv. Geophys.*, **31**, 107–129, <https://doi.org/10.1007/s10712-009-9079-x>.
- Wapler, K., and P. James, 2015: Thunderstorm occurrence and characteristics in Central Europe under different synoptic conditions. *Atmos. Res.*, **158–159**, 231–244, <https://doi.org/10.1016/j.atmosres.2014.07.011>.
- Weingart, N., 2018: Deep learning based error correction of numerical weather prediction in Switzerland. M.S. thesis, Systems Group, Department of Computer Science, ETH Zurich, 67 pp.
- Wilson, J. W., N. A. Crook, C. K. Mueller, J. Sun, and M. Dixon, 1998: Nowcasting thunderstorms: A status report. *Bull. Amer. Meteor. Soc.*, **79**, 2079–2099, [https://doi.org/10.1175/1520-0477\(1998\)079<2079:NTASR>2.0.CO;2](https://doi.org/10.1175/1520-0477(1998)079<2079:NTASR>2.0.CO;2).
- Zawadzki, I., 1973: Statistical properties of precipitation patterns. *J. Appl. Meteor.*, **12**, 459–472, [https://doi.org/10.1175/1520-0450\(1973\)012<0459:SPOPP>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0459:SPOPP>2.0.CO;2).
- Zeder, J., U. Hamann, D. Nerini, L. Foresti, L. Clementi, A. Hering, and U. Germann, 2018: Comparison of thunderstorm characteristics as seen by SEVIRI and radar regarding lightning and hail initiation. *EUMETSAT Meteorological Satellite Conf.*, Tallinn, Estonia, MeteoSwiss, <https://doi.org/10.13140/RG.2.2.23453.77285>.