In Search of Micromodels: Black-box Evaluation of Spatio-temporal Models in Normal and Extreme Conditions

Ivana Nižetić Kosović, Toni Mastelić, Domina Sokol, and Diana Škurić Kuražić

Original scientific article

***Abstract*—Spatio-temporal modelling is an emerging research area due to the increasing availability of sensor data collected across space and time. The models are build either with a model-driven or data-driven approach. The former often results in complex monolith models that are not suitable for lightweight Edge deployment. The latter requires a vast amount of data and may not provide an overall good performance. Consequently, the data-driven approach is being used to substitute only parts of model-driven outputs, by creating micromodels that tackle specific scenarios. The main contribution of this paper is a definition and demonstration of the process for finding such scenarios for which a spatio-temporal model could be improved or replaced by a micromodel and deployed on Edge. The process is demonstrated on an example of a Numerical Weather Prediction model (NWP), namely its outputs of temperature and precipitation. NWP is evaluated using black-box testing considering the specificity of spatial and temporal components, in both normal and extreme conditions. The novelty of this process is its ability to highlight weaknesses of the existing expert models and suggest scenarios in which the models can be improved and deployed on the Edge.**

***Index Terms*—Model validation, data-driven approach, extreme events, Numerical Weather Prediction (NWP), Machine learning (ML).**

I. Introduction

In the era of huge automatic data collection using variety of sensors, spatio-temporal data is gaining more attention [1] [2]. The specificity of such data comprises both spatial and temporal attributes. Spatio-temporal data is the base for many applications, such as transportation, climate, Earth science, etc. Related models represent the temporal change of spatial objects or phenomena over time [3], making them excellent candidates for the Edge computing [4] paradigm. Consequently, the data collected at certain locations can be analysed, modelled, and used at the same location without sending a vast amount of data to the Cloud and back.

There are two approaches when creating the models: a model-driven and data-driven approach [5]. The model-driven approach (also called a knowledge-based or physical models) is based on domain knowledge. The model simulates the behaviour describing the laws of the domain (e.g. kinematics, dynamics, thermodynamics, etc.). The data-driven approach, enabled by machine learning [6] [7], exploits data to automatically learn the patterns that occur in a system or a process.

A. Problem Statement

The model-driven approach relies on a deep understanding of a system or a process. This leads to the creation of complex monolith models with high modelling and execution costs. Such models commonly cannot leverage the advantages of the Edge paradigm. Data-driven models are easier to build and use. Although data-driven models need a vast amount of data to learn variations in data, they can easily be divided into smaller models, referred to as micromodels [8]. This is especially applicable considering the spatio-temporal nature of the data. Such micromodels are specialized for a particular task or a particular piece of data [8]. They might not perform well on general problems, but can be very effective when solving specific problems. Finally, micromodels can be small enough to run on Edge devices.

To find the potential scenarios for micromodels, the original model should be properly evaluated in chosen scenarios, considering both the spatial and temporal properties of the model. Moreover, the model should be additionally evaluated for extreme events. Extreme events are one of the major concerns in Earth sciences [9] which highly utilize spatio-temporal data. Internal mechanisms of the system, described by model-driven models, are known. Accordingly, it is common that model-driven models are evaluated using white-box testing (similar to verification) [10]. Such an approach is well suited for detecting defects in the internal structure of the model. However, it might not be good in revealing scenarios where micromodels could be applied to improve performance.

B. Related Work

At the 2021 Artificial Intelligence Summit in New York, the presentation on machine learning challenges with micromodelling broke the myth that having more data is always better [11]. In many cases, modelling specific task on small amounts of data can yield better results. Lately, many companies have embraced the concept of micromodelling. Microsoft [12] is finding micromodels scalable for model training and training

Manuscript received July 27, 2022; revised August 29, 2022. Date of publication September 26, 2022. Date of current version September 26, 2022. The associate editor prof. Claudia Canali has been coordinating the review of this manuscript and approved it for publication.

I. Nižetić Kosović, T. Mastelić and D. Škurić Kuraži are with the Research Department of Ericsson Nikola Tesla, Croatia (e-mail: ivana.nizetic.kosovic@ericsson.com). D. Sokol is with the Faculty of Science, Split, Croatia.

Digital Object Identifier (DOI): 10.24138/jcomss-2022-0092

them in parallel. Companies claim that micromodels ease their businesses and produce better predictions [8] [13] [14]. In [15] the idea is to move away from big comprehensive models to multiple micromodels trained to solve only a fragment of a complex system in the software engineering process.

An example of a micromodel use is [16], in which the authors built the task-specific models for mental health. In [12] authors present a set of simpler models for optimizing big data workloads. There are many other examples of using micromodels that replace part of the complex domain expert system. One of the examples is the replacement of a pyranometer (which is the important part of the smart building managing systems) with the low-cost set of sensors and software [17]. Another example is an early estimation of seawater quality, replacing a complex microbiological analysis of the seawater sample [18]. In many cases, they are not called "micromodels" explicitly, but the concept remains the same.

In the above mentioned research, the scenarios in which micromodels are deployed are already known, recognized by the domain expert. In cases where there is no prior knowledge on potential scenarios for micromodels, the model itself should be evaluated. In that way, one will be able to pinpoint the task candidates for micromodels.

In the examples of Numerical Weather Prediction (NWP) reports [19] [20], the models are evaluated using white-box testing. In [19] the verification is performed by presenting the model change over the years. The model is compared to other weather forecast models, against various model configurations. Similarly, in [20] the authors evaluate the ALADIN model. They claim that the ALADIN model forecasts highly depend on the resolution of the model, data used by the model, and initial parameters of the model. They also suggest that the model is not suitable for weather events characteristics typical for Croatia. They do not provide the error for forecasting parameters considering the spatial or temporal variability, which may reveal niches where micromodels could be applied.

Although pointing out model weaknesses, these reports aim to enhance the internal structure of such monolith models. Such models still may continue to fail in certain scenarios and still may not be suitable for the forthcoming Edge deployment.

C. Contributions of the Paper

Our approach is to use black-box testing [10] for model evaluation. Unlike white-box testing, black-box testing pretends that the internal mechanisms of the system are unknown. Accordingly, the model-driven spatio-temporal model is thus evaluated considering the specificity of spatial and temporal components, rather than its internals. The model is evaluated in both normal and extreme conditions. Following this approach, we can find scenarios in which the original complex spatio-temporal models could be improved, preferably replaced by a micromodel, and deployed on Edge.

To demonstrate this process, the Numerical Weather Prediction (NWP) model is taken as a use case, representing the model-driven spatio-temporal model. The great potential of deep learning and machine learning models has been recently recognized as the aid to the weather prediction process [21] [22]. These models have the ability to discover hidden patterns that are not part of the traditional NWPs [23]. Furthermore, data-driven models could be useful in scenarios in which the existing model fails to model the phenomena well. These scenarios are the perfect candidates for micromodels. The process of finding such scenarios is demonstrated on the NWP model outputs of two parameters: temperature and precipitation. The outputs are available on the webpage of the public service (Croatian Meteorological and Hydrological Service - DHMZ [24]). They are compared with the measurements from standard meteorological stations taken as the ground truth. The measurements are available on the webpage of the Reliable Prognosis service [25]. The analysis is performed on the 19 months data for six locations (cities) in Croatia.

Accordingly, the contributions of this paper include:

- the definition of the approach of searching for scenarios in which the original spatio-temporal model could be improved.
- the demonstration of the above approach on the NWP model.

D. Structure of the Paper

The rest of the paper is organized as follows. In the next section, the overview of weather forecasting models is given. Section III consists of the descriptions and insights into the datasets used in the analysis. The explanation of our analysis approach is given in Section IV. The Section V presents the results of the analysis. This section is divided into three subsections. The first subsection shows the behaviour of the NWP model compared to the baseline models. The second subsection demonstrates the analysis considering different time and space aggregations. The third subsection shows the analysis of forecasting considering extreme weather conditions. Finally, the discussion section summarizes the results and the conclusion section proposes future research directions.

II. Background

Three meteorological expert models are evaluated in this study. The first is the Numerical Weather Prediction (NWP) model which serves as the main model being examined while searching for potential micromodels. Two baseline models serve for evaluating the performance of NWP, which are also used to indicate scenarios where micromodels may be beneficial.

A. Numerical Weather Prediction

The most common approach to forecasting future weather (e.g., upper and surface air pressure, temperature, wind speed, relative humidity, etc.) is NWP modelling. The concept of NWP is to solve a set of partial differential equations that simulate physical laws of atmosphere variables [26]. To solve the equations, initial and boundary conditions are used. For initial conditions (the estimation of the present state of the atmosphere), weather observations from standard meteorological stations, atmospheric soundings, and remote sensors (satellites) are used. Boundary conditions, which define the

atmosphere's state and the domains' edges, depend on the region which is being simulated. For example, the whole Earth will be taken in global models, the continent or part of the continent for regional models, etc.

The data used for initial conditions are transformed to form grid-shaped data. Global models have a typical horizontal resolution of around 100 km and vertical resolution of 25–50 hectopascal [27]. The prediction values are not taken solely by running NWP, rather the data are additionally processed during the phases of data assimilation and post-processing. It can be seen that the NWP models become better and better through the years [19]. Modern NWPs exploit an ensemble of forecasts built on slightly different initial conditions and take ensemble mean to produce a single forecast [23] [27].

Croatian Meteorological and Hydrological Service (DHMZ) provides forecast outputs based on two NWP models: the regional Aire Limitée Adaptation dynamique Développement InterNational (ALADIN) [28] and the global European Centre for Medium-Range Weather Forecasts (ECMWF) [29]. ALADIN model produces forecasts four times per day (00 UTC, 06 UTC, 12 UTC, and 18 UTC) up to 72 hours ahead. The horizontal resolution of this regional model is 8 km with narrowed areas in Croatia where dynamic adaptation enables forecasts with a horizontal resolution of 2 km. This model has been continuously developed by the experts [30]. ECMWF's model, as a global model, is developed and executed in ECMWF headquarters in Reading, UK. Their forecasts and other data are directly available to their Member and Co-operating States [31]. This model produces forecasts on a wide temporal range: from medium (up to 15 days), extended (up to 42 days) to long range (up to several months). For the purposes of this article, we focus on ECMWF's medium-range forecasts that are available twice a day (00 UTC and 12 UTC) up to 7 days ahead, with a horizontal resolution of 16 km [29].

In this paper, the aim is to evaluate the final outputs of the NWP model, which are available to the public, Therefore, the complexity of decisions being taken during the process is neglected during our evaluation.

B. Baseline Models

For the sake of comparison, two simple models are also evaluated: the persistence model and the model based on climatology. They are considered important baselines in weather forecasting. They are often used as a reference for determining the performance of the observed model, in our case the NWP model [32].

On the one hand, the persistence model assumes that the next forecast value is the same as the current value. In terms of the weather forecast, the persistence model will predict the same values for all future timestamps. On the other hand, climatology refers to the average values of observed phenomena for a certain space and time over a certain climate period (usually 30 years). Usually, averaged values in climatology are based on a daily or monthly time scale. In this paper, to be able to fairly compare the NWP model with climatology, we constructed the model based on climatology. It works as follows: the average of 15 years of data values for certain phenomena is taken for each 3-hour periods. The model based on climatology will predict the values for the timestamps ahead as the average value of past years. This is done regardless of the year at the time of of forecasting.

III. Materials

Two datasets are used to conduct this study. One is the dataset based on the Croatian Meteorological and Hydrological Service (DHMZ) weather forecasts, and the other one includes the actual meteorological measurements acquired from the Reliable Prognosis webpage [25]. The datasets are combined, filtered, and prepared for analysis as described in more detail at the end of this section.

A. Forecasts Dataset

The dataset for 7-day forecasts contains the data from 4th February 2020 to 21st September 2021 for 85 locations in Croatia. The final forecast outputs that comprise two models: ALADIN [20] (from the 1st to the 3rd day of forecast) and ECMWF [33] (from the 4th to the 7th day of forecast). The forecasted values are given in the steps of three hours. This output is updated by the DHMZ multiple times in a day, producing multiple outputs for the same forecasted time. We took only the first run of the forecast while discarding the updates. The original dataset contains several weather parameters. We chose temperature and precipitation for further analysis.

The dataset contains the following information:

- forecast time (date and time - CET)
- location for which the forecast is made (city name)
- predicted precipitation (in mm/3h)
- predicted temperature (in degrees Celsius)
- time when the forecast was generated by the model (date and time - CET)
- model run (hour of the day - 00, 06, 12 or 18 UTC)

Date and time data are converted from CET to UTC, with regards to daylight savings. As forecasts are given for every three hours, this produces forecast times of 00:00, 03:00, 06:00, etc.

The last day of the forecast (the 7th day) has forecast times every six hours. Since the exact hours vary, the 7th day is removed from the analysis for all locations.

There are around 3.8% of missing forecasts, mostly in May 2020. Consequently, some of the timestamps have too few forecast values. Timestamps with less than 50 % of forecast values are removed from the dataset.

B. Measurements Dataset

The actual measurements of temperature and precipitation for 100 locations in Croatia are downloaded from the archive of the Reliable Prognosis webpage [25] for the period from 1st January 2005 to 31st December 2021. The dataset is used twofold: to produce the model based on climatology (from the year 2005 to 2019) and to evaluate the NWP and baseline models (from the year 2020 to 2021). The dataset includes more variables than the forecast dataset from III-A.

Fig. 1. Locations in Croatia selected for the analysis

For further analysis, only those needed for the paper are selected:

- local time (UTC + 1h)
- temperature (in degrees Celsius)
- precipitation (in mm/6h and mm/12h)

Precipitation was given in millimeters per different time intervals; mainly during the last 6 and 12 hours. As these time intervals were almost always regularly exchanged one after another, it was possible to deduce the amount of precipitation every 6 hours for all precipitation records. To match the three-hourly forecasts, the precipitation amounts were then averaged and divided into 3-hour bins. Several precipitation outliers were removed as they were larger than the other precipitation values by two orders of magnitude. This produced a loss of only 7 data points, mainly in 2009.

C. Analysis Datasets

The analysis is performed on two continuous variables: temperature and precipitation. Out of all locations, six cities (Dubrovnik, Split, Zadar, Rijeka, Zagreb, and Osijek - see Figure 1) have been considered for two reasons. First, they are the largest cities in different regions of Croatia. Therefore, they cover a variety of climates. Four of them are coastal cities (Dubrovnik, Split, Zadar, Rijeka) and two are continental cities (Zagreb, Osijek). The other reason is that there is a large volume of historical weather data available for these cities.

Due to the missing data, there is some discrepancy between the number of timestamps that are preserved for each location. To be able to fairly compare locations, those timestamps that are present for each location are preserved. The final dataset consists of 4081 forecast timestamps of a certain parameter for each of the six locations. Characteristics of the selected dataset by locations are depicted in Figure 2, along with the climatology data.

Figure 2a) depicts temperatures and Figure 2b) depicts precipitation. The evaluation dataset is represented in orange and the climatology dataset in blue. Wide boxes represent 50% of data. Thin vertical lines at the beginning and at the end of the boxes are the minimal and the maximal values. The medians are marked with thick vertical lines. The precipitation values for measurements with precipitation higher than zero (”no precipitation” values) are excluded from the right hand side of Figure 2b). Most of the measurements have a small amount of precipitation (below 1mm/3h as can be seen from the right hand side of the graph). Therefore, the values are presented on a logarithmic scale of the original values. The percentage of values taken on the right-hand side of the graph is presented on the left-hand side. Around 75% to 82% data are with the precipitation value zero, meaning there is no precipitation.

IV. Methods

In this study, we utilize the black-box evaluation approach, by examining model outcome errors considering spatial and temporal variability. Consequently, the evaluation metric used across the evaluation process in this paper is the root mean squared error (RMSE) [34] given by the formula:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(v_i - \widehat{v}_i)^2}{n}}$$

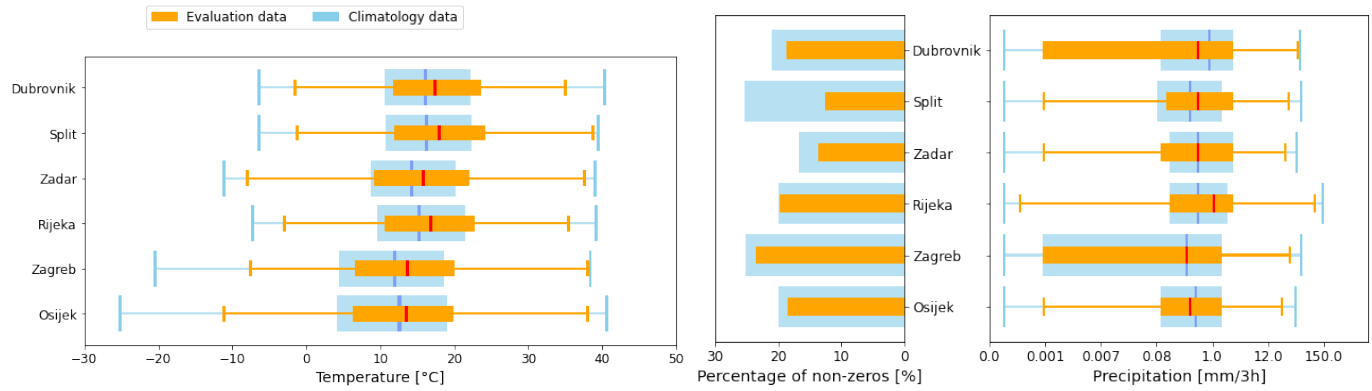
where v_i is an actual measurement value and $\widehat{v}_i$ is the forecasting value for i-th observation.

The errors of the model are shown as a function of the forecast lead time with a 3-hour resolution. This means that the errors depict the forecast 3, 6, 9, 12, .., 144 hours ahead (see Figure 3 - a row represents a future timestamp, and a column represents a lead time).

In our search for micromodels, we perform three groups of evaluation tests, each of which consists of several sub-evaluations, namely:

A. *Model comparison.* This analysis aims to compare different weather prediction models.
   1. *Comparison of NWP with baseline models.* NWP model is compared with two baseline models: the persistence model and the model based on climatology. The evaluation is presented considering a change of RMSE when changing lead time for all locations together. Residuals for each model are analyzed for each location separately.
   2. *Comparison of NWP model runs.* Different NWP models (models with different runs) are compared to each other, for each location separately.

B. *NWP model performance considering space and time differences.* This analysis answers how the NWP model differs considering different subsets of data regarding space and time.
   1. *Location differences.* Forecasts for each location are considered separately.
   2. *Seasonal differences.* Seasonal forecasts (winter, spring, summer, and autumn) are compared for each location.
   3. *Differences in time of the day.* Forecasts for day and night are compared for each location.



(a) Temperature (b) Precipitation

Fig. 2. Descriptive analysis for (a) temperature and (b) precipitation for six locations in Croatia for a 15-years climatology and evaluation dataset.

time	3	6	9	12	15	18	21	24
2020-02-08 00:00:00	-	3.0	-	2.0	-	2.0	-	2.0
2020-02-08 03:00:00	3.0	-	2.0	-	2.0	-	1.0	-
2020-02-08 06:00:00	-	2.0	-	1.0	-	1.0	-	1.0
2020-02-08 09:00:00	6.0	-	5.0	-	5.0	-	5.0	-
2020-02-08 12:00:00	-	10.0	-	10.0	-	10.0	-	10.0
2020-02-08 15:00:00	11.0	-	10.0	-	10.0	-	10.0	-
2020-02-08 18:00:00	-	5.0	-	5.0	-	4.0	-	4.0
2020-02-08 21:00:00	4.0	-	4.0	-	3.0	-	3.0	-

Fig. 3. Example of NWP predictions for one location for one day considering time lags.

C. *NWP model performance for extreme conditions.* This analysis aims to compare the model results on a subset of extreme conditions, for each location separately.
Two types of anomalies are considered in the analysis: value anomalies and delta anomalies. Value anomalies are a type of extreme conditions which describe deviations from the expected values (above or below a certain percentile). Delta anomalies are a type of extreme conditions which describe high changes in values (below or above a certain percentile).

Results of the evaluation of the NWP model are presented in the next section (Results). The enumeration in the next section follows the enumeration introduced here. The suggestions of the scenarios in which the micromodels could be deployed are given at the end of the subsections.

V. Results

A. Model Comparison

In this section, we compare the NWP model with baseline models (both the persistence model and the model based on climatology) and we compare different NWP models (models with different runs).

1) Comparison of NWP Model with Baseline Models: The comparison of the baseline models and the NWP model for temperature forecast for all six locations in Croatia is given in Figure 4a. The error for the NWP model (red) increases as the leap time increases (as expected) from around 2°C for 3 hours ahead up to around 3°C for 6 days ahead. The model based on climatology (blue) is constant - its ”forecast” is always the same, regardless of the lead time. However, small alternation can be seen in the Figure, due to the different underlying subsets of the evaluation dataset. For example, the timestamp at 15:00 UTC has forecasts of 3, 9, 15 hours, etc. lead time, but no forecast of 6, 12, 18 hours, etc. lead time, since the forecast is generated every six hours (see Figure 3). The model’s error is around 3.8°C.

The persistence model (yellow) shows daily periodicity in error, showing smaller errors every 24 hours. This is expected since the model ”forecasts” the future values to be the same as the present values. The temperature value for the next day at the same time of the day is likely to be predicted with a smaller error than the value for the other part of the day. The error of the persistence model by lead time goes from 3°C to 7°C.

The comparison of the baseline models and the NWP model for precipitation for all six locations in Croatia is given in Figure 4b. The error for the NWP model (red) and the error for the model based on climatology (blue) are quite similar (around 2 mm/3h). The largest error is achieved for the persistence model (yellow - around 2.5 mm/3h). The errors for each of the models are similar regardless of the lead time. The error is calculated for all timestamps in the evaluation dataset. Around 80% of the data in the dataset are 0 mm, meaning ’No precipitation’ (as it is shown in section III-B). Most of the precipitation values are below 1 mm/3h, with several events of heavy rain (with a maximum value of 150 mm/3h).

An additional analysis is performed to compare residuals of the forecasts for different locations and different models (see Figure 5). The NWP model has the smallest spread, while the persistence model has the largest spread. The NWP tends to slightly underestimate locations: Dubrovnik, Split, and Rijeka; and slightly overestimate Osijek. The model based on climatology has the mean closest to zero.

A comparison of precipitation residuals of the forecasts for different locations and different models is given in Figure 6. The first subfigure 6a) illustrates how good the models are in forecasting the precipitation as if it were a classification problem (true - precipitation, false - no precipitation). Both NWP and persistence models are good in predicting no precip-

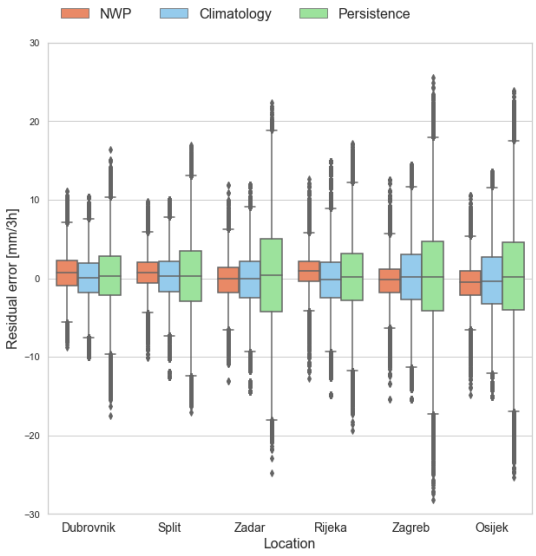


Fig. 5. Residual error spread for temperature, for each location and models: NWP, climatology, and persistence.

itation (TN). For all locations except Rijeka, the best model for predicting the presence of precipitation is the model based on climatology. Rijeka is a location with high amounts of precipitation - see Figure 2. The best model for Rijeka in predicting the presence of precipitation is NWP (TP). The model based on climatology also tends to slightly underestimate the amount of precipitation. Subfigure 6b) shows the spread of residuals of models in forecasting the exact amount of precipitation, in TP situations. The lowest spread is achieved by the model based on climatology. The NWP and the persistence model show almost the same behaviour. In subfigure 6c) the spread of residuals of models in FN situations is shown. This includes cases when models predicted no precipitation when the actual precipitation was above zero. The lowest spread is achieved by the NWP model. Finally, subfigure 6d) presents the spread of residuals in FP situations. This includes cases when the models predicted precipitation above zero when the actual precipitation was zero. Significantly lower spread is achieved by the model based on climatology.

2) Comparison of NWP Model Runs: The evaluation dataset used in this paper consists of forecasts generated by the NWP models four times per day (run = 00, 06, 12, and 18 UTC), as described in section III-A.

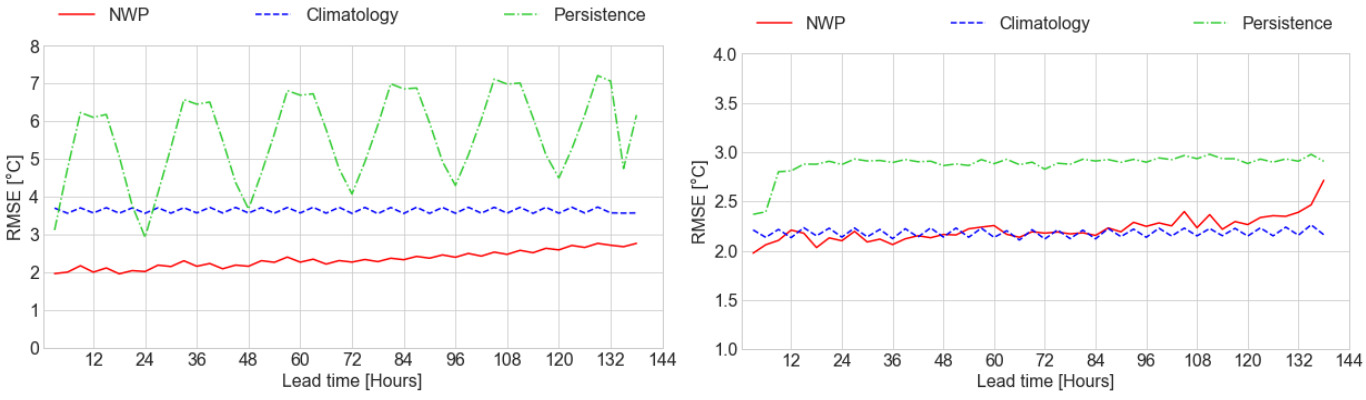
Both the temperature errors and precipitation errors by model runs are shown in Figure 7 The error behaviour of each run is periodical. For temperature (Figure 7a)), lower errors are achieved for the middle-of-the-day forecasts. For example, a temperature forecast with run 00 UTC will give the lowest error 12 hours ahead, the one with run 12 UTC, 24 hours ahead, etc. For precipitation (Figure 7b)), lower errors are achieved for the afternoon forecasts. For example, precipitation forecast with run 00 UTC will give the lowest error 18 hours ahead, the one with run 12 UTC, 6 hours ahead, etc.

B. NWP Model Performance, considering Space and Time Differences

In this section, we evaluate the NWP model considering different subsets of data regarding space and time.

1) Location Differences: To check for potential location differences, NWP model temperature error is shown separately for each location in Figure 8a. The graph shows the behaviour of forecast errors six days ahead by location. There is a difference in forecast error behaviour for the first three days compared to the next three days of the NWP forecast. This is the consequence of different NWP models used to generate forecasts: the first three days are forecasts from ALADIN and the next three days are forecasts from the ECMWF's model. Furthermore, there are differences in errors regarding locations. The behaviour most similar to the one in figures 4 and also the most expected behaviour is for location Split (dark green) - the error slightly increases as the lead time increases. The same behaviour for the first three days is shown for Dubrovnik (dark blue), but its error unexpectedly increases after the third day (from 2°C to 3°C). Even more unexpected behaviour can be seen for the set of locations: Osijek, Zagreb, Zadar, Dubrovnik, and Rijeka, for which the forecast error is higher for the first three days than for the next three days. For example, for Osijek, the forecast error increases up to even 3°C in the first three days and then decreases to around 2°C. For some locations, an error for the first three days forecast shows daily periodicity.

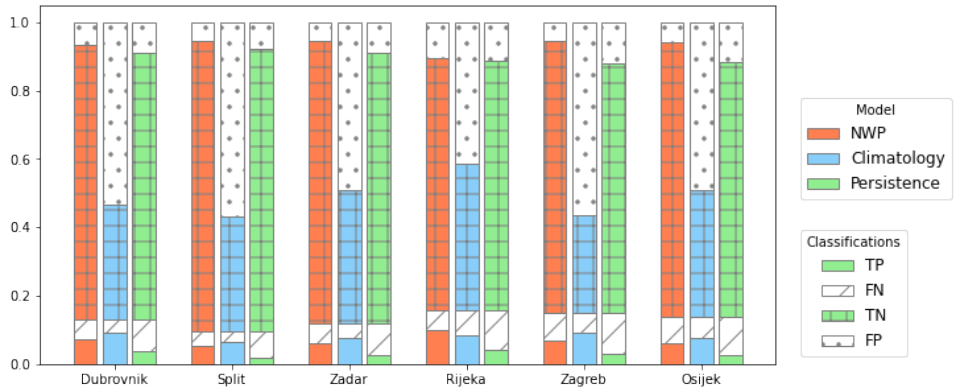
NWP model precipitation error is shown separately for each location in Figure 8b. The largest error (4 mm/3h) is



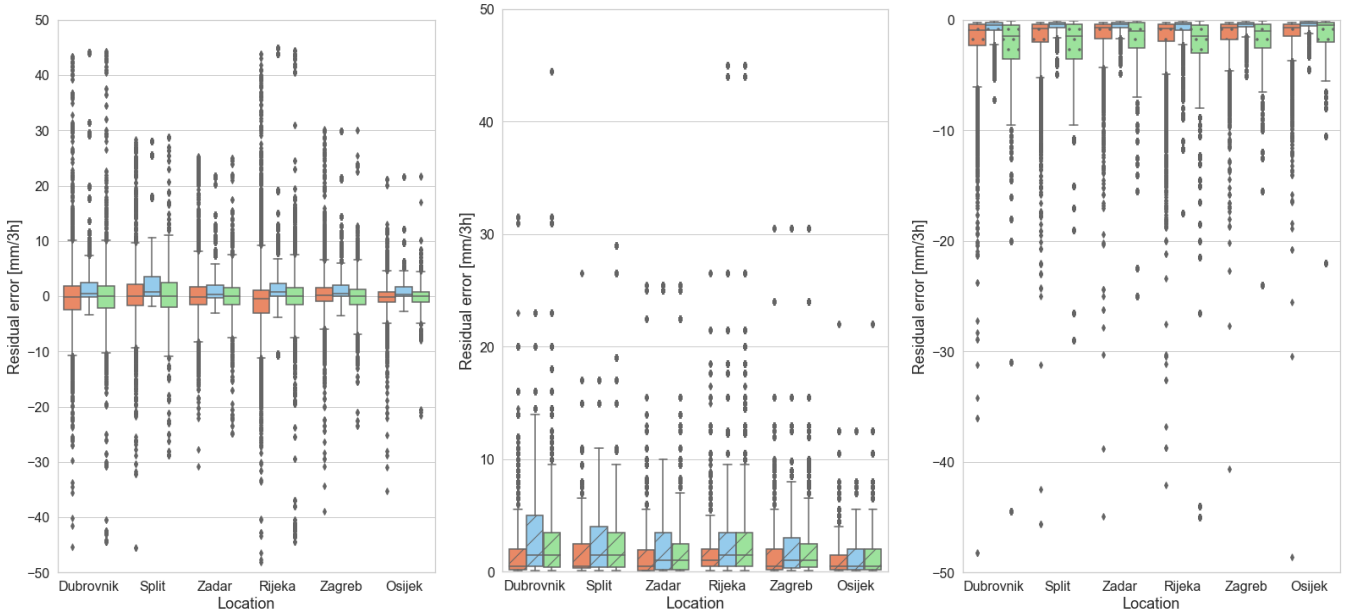
(a) Temperature

(b) Precipitation

Fig. 4. Error (RMSE) of NWP and baseline models for (a) temperature and (b) precipitation.

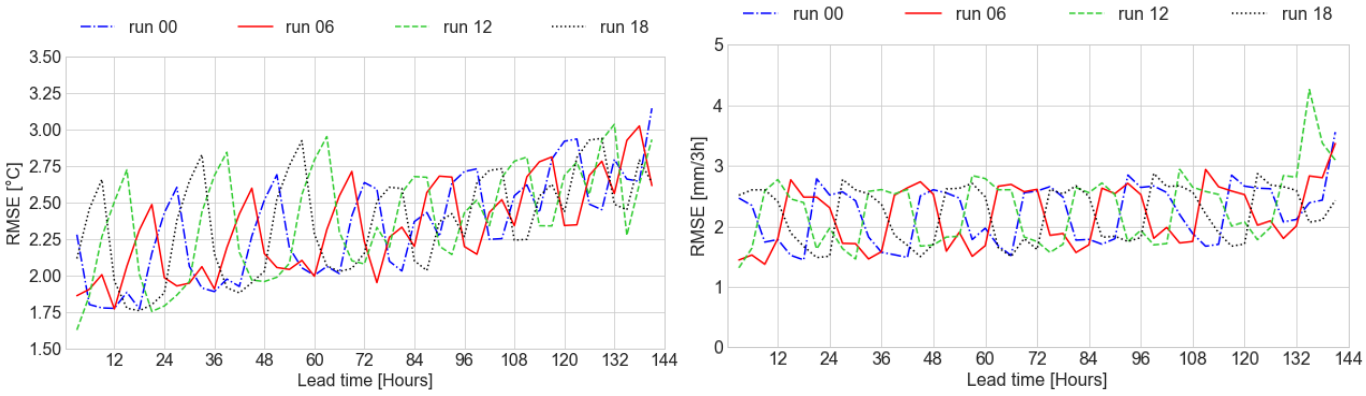


(a) Precipitation classification error



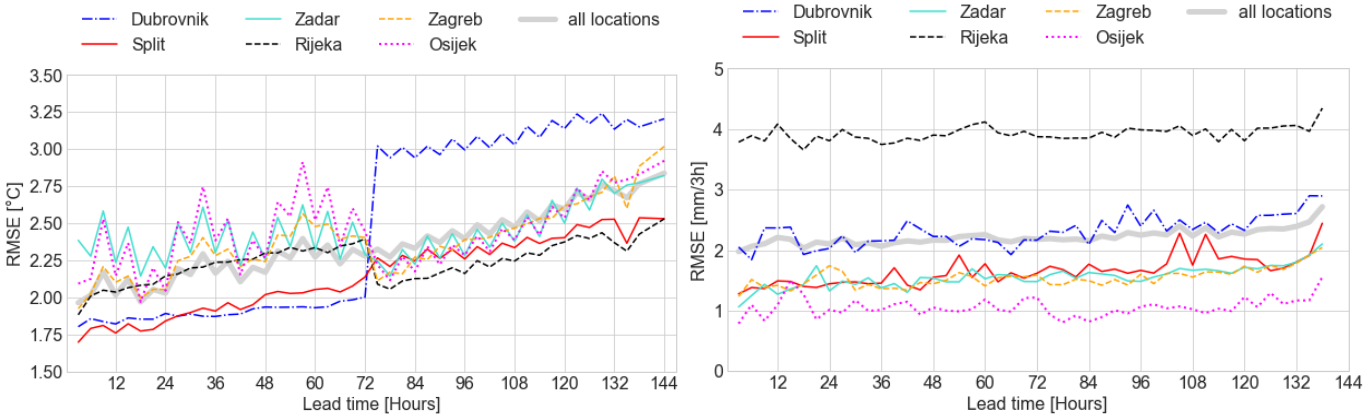
(b) Precipitation error for TP (c) Precipitation error for FN (d) Precipitation error for FP

Fig. 6. Precipitation error spread considering true positives, false negatives, true negatives, and false positives. (a) Percentage of precipitation classification errors; Residual error spread for precipitation (b) for true positives - TP, (c) for false negatives - FN, (d) for false positives - FP, for each location and model: NWP, climatology, and persistence.



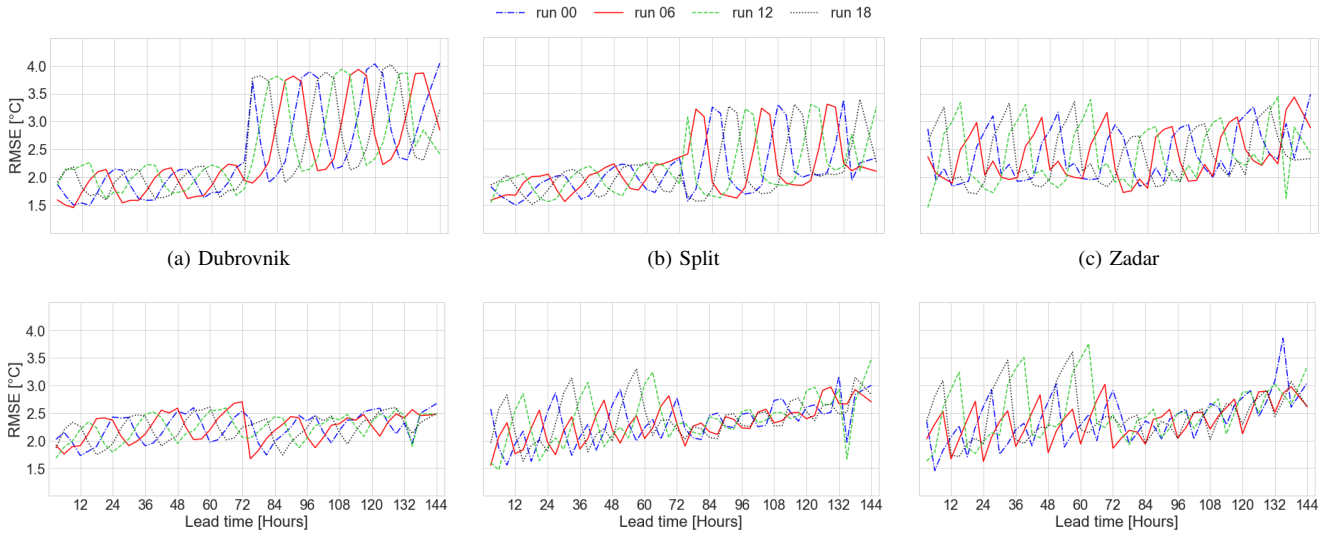
(a) Temperature (b) Precipitation

Fig. 7. Error (RMSE) of NWP model by runs (blue - run 00, red - run 06, green - run12, grey - run 18) for (a) temperature and (b) precipitation.



(a) Temperature (b) Precipitation

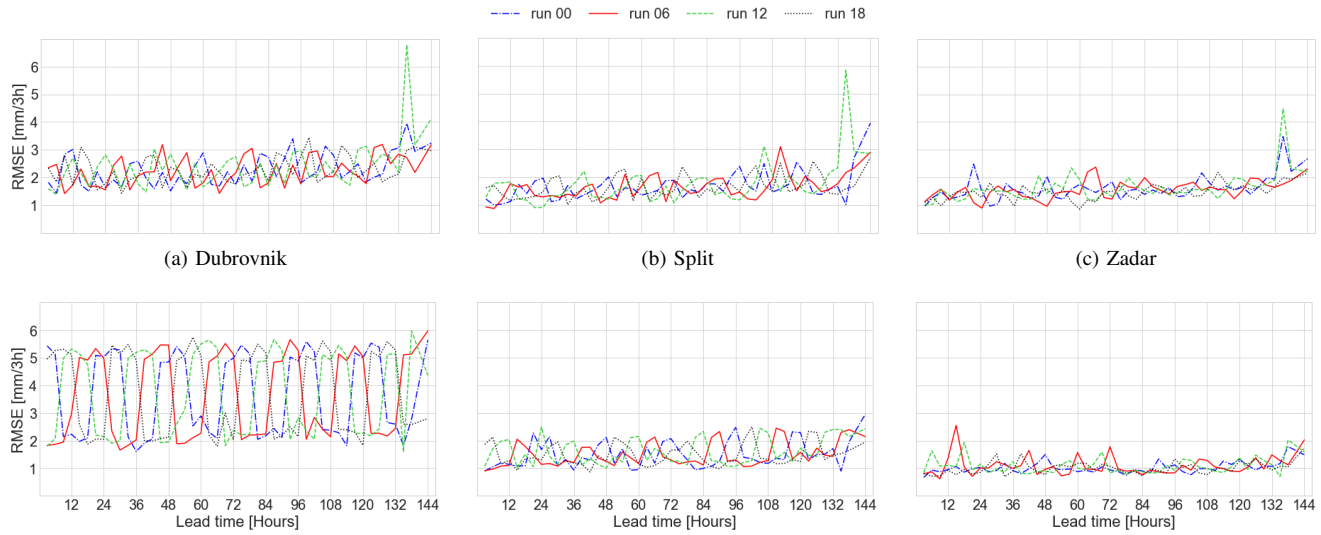
Fig. 8. Error (RMSE) of NWP model for different locations for (a) temperature and (b) precipitation.



(a) Dubrovnik (b) Split (c) Zadar

(d) Rijeka (e) Zagreb (f) Osijek

Fig. 9. Temperature error for different locations by runs



(a) Dubrovnik (b) Split (c) Zadar

(d) Rijeka (e) Zagreb (f) Osijek

Fig. 10. Precipitation error for different locations by runs

achieved for Rijeka, while the smallest error (1 mm/3h) is achieved for Osijek. Rijeka has a significantly larger error than other locations. Rijeka has the largest amounts of precipitation (depicted in section III-B). This could be the reason for large errors for that location.

Additionally, we also evaluate different NWP models (by runs) for each location separately. As previously described, the error behaviour of each run is periodical. This can be seen for each location as well, both for temperature and the precipitation (Figure 9 and Figure 10). The temperature graphs (Figure 9) show the difference in the forecast error amplitude for the first three and the next three days, which is not consistent for each location. For example, errors in temperature forecasts for Split and Dubrovnik are more stable in the first three days than in the next three days. The opposite behaviour is seen for Zagreb and Osijek, while Rijeka has the most stable and Zadar has the most unstable error behaviour for all lead times. The graphs of the precipitation error for each location by model runs (Figure 10) show the most unstable results for Rijeka. For Rijeka, the error varies from 2 mm/3h up to 5 mm/3h. The lowest errors are achieved for Zadar and Osijek (around 1 mm/3h).

2) Seasonal Differences: The comparison of errors regarding the season (winter, spring, summer, and autumn) is given in Figure 11 for each location. The behaviour again differs for different locations. For example, the temperature error for Split is the smallest in the summer (1.5°C-2°C) and the largest in the winter (2°C-2.5°C), in the first three days. In the next three days, the behaviour is the opposite - the error is the smallest for the winter and the autumn (1.5°C-2.5°C) and the largest for the summer (2.5°C-3°C). For Zagreb, Rijeka, and Osijek, the difference is present in the first three days, while quite similar results are gained for each season for the next three days. For Zagreb and Osijek, the largest errors in the first days are achieved for the summer, while for Rijeka, the largest errors are achieved for the winter. Dubrovnik has the most stable results, although the jump in error after the third day is visible for each season. The largest jump is shown for autumn (almost 2°C difference between the first three days and the next three days).

The comparison of precipitation errors in different seasons (winter, spring, summer, and autumn) is given in Figure 12 for each location. The behaviour again differs for different locations. Split, Dubrovnik and Zadar achieve the largest error for the autumn (around 3 mm/3h, 4 mm/3h, and 4 mm/3h respectively). These locations also achieved the lowest error for the spring and the summer (around 1 mm/3h). Zagreb and Osijek have more similar results for each season, although slightly larger errors are achieved for the summer (around 2 mm/3h and 1 mm/3h respectively). The smallest errors are achieved for Osijek for the summer (around 0.5 mm/3h).

3) Differences in Time of the Day: The comparison of errors regarding the time of the day (day and night) is given in Figure 13. The error is in general smaller for day values. The exception is Split, which has a smaller error for the night after the third day.

The comparison of precipitation errors regarding the time of the day (day and night) is given in Figure 14. The error is quite similar for day and night for all locations, except for Rijeka.

C. NWP Model Performance for Extreme Conditions

In this section, the NWP model is evaluated only with regard to extreme values, and the results are compared to the results on the whole dataset. Two types of extremes are taken into the analysis: value anomalies and delta anomalies.

To subset the value anomalies, the thresholds are taken from the Climate Explorer [35] for each of the six locations (based on 30 years of data from 1990 to 2019). The temperatures above the 95th percentile of the maximum daily temperatures and the values below the 5th percentile of the minimum daily temperatures are considered as value extremes. Since we are interested only in extremely high (not extremely low) precipitation, the precipitation above the 90th percentile of average daily precipitation is considered as value extreme. Percentiles are calculated based on climatology, and the evaluation is performed on the same dataset as in the previous sections.

For delta anomalies, all consecutive changes in values in the dataset consisting of 5 years of data (from 2015 to 2019 - from (III-B) are calculated. The 5th and the 95th percentile are calculated for each location. Changes in the evaluation dataset that are below the 5th or above the 95th percentile are considered delta extremes for both temperature and precipitation. The amount of both extreme types filtered in this procedure is given in Table I.

Temperature extremes for each location are depicted in Figure 15. The RMSE is higher for both conditions that belong to anomalies than the RMSE on the whole dataset. In most cases, worse results are achieved for value anomalies.

Precipitation extremes for each location are depicted in Figure 16. For all locations, the RMSE is significantly higher for events that belong to value anomalies. The largest error is achieved for Rijeka (around 25mm/3h). Compared to the results for temperatures, the NWP is worse in predicting extreme precipitation.

VI. Discussion

In the previous sections, the analysis indicates many opportunities for model improvement. When comparing models,

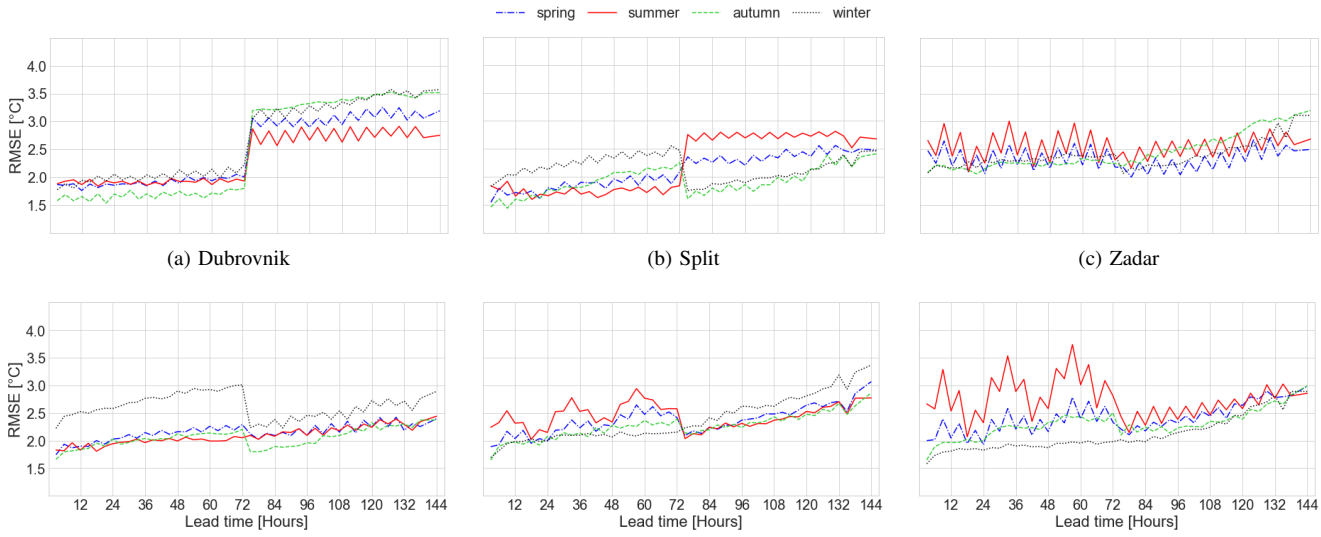
TABLE I
NUMBER OF VALUE ANOMALIES AND DELTA ANOMALIES ON EVALUATION DATASET FOR TEMPERATURE

	Temperature		Precipitation	
Location	Nr. of value anomalies	Nr. of delta anomalies	Nr. of value anomalies	Nr. of delta anomalies
Dubrovnik	125	176	77	356
Split	151	175	89	291
Zadar	584	181	89	410
Rijeka	61	166	72	396
Zagreb	283	173	71	437
Osijek	88	173	67	410

the baseline model based on climatology can yield forecasting almost as accurate as the NWP model considering temperature (Figure 5). There is room for improvement considering the precipitation error, especially when the precipitation is present. The model based on climatology shows better results when predicting the presence of precipitation. The same model shows the worst results when predicting no precipitation (Figure 6). That indicates that the combination of those two models can yield better overall accuracy. The benefit of using the model based on climatology instead of the NWP model is its speed and ease of deployment. The model with different runs shows very different but regular forecasting errors (Figure 7), leaving room for improvement. Combining the forecasting from different runs into one forecast could result in better overall accuracy, for both temperature and precipitation. For example, using run 06 for the first 6 hours, run 00 for the next 6 hours, etc. for precipitation will yield higher overall accuracy.

The analysis shows that the development of models for specific locations could lead to higher accuracy (Figure 8). This is especially visible in some locations. An example is Dubrovnik which, compared to other locations, has an extremely large temperature error from the third to the sixth day of the forecast. Similarly, compared to other locations, Rijeka has an extremely large precipitation error.

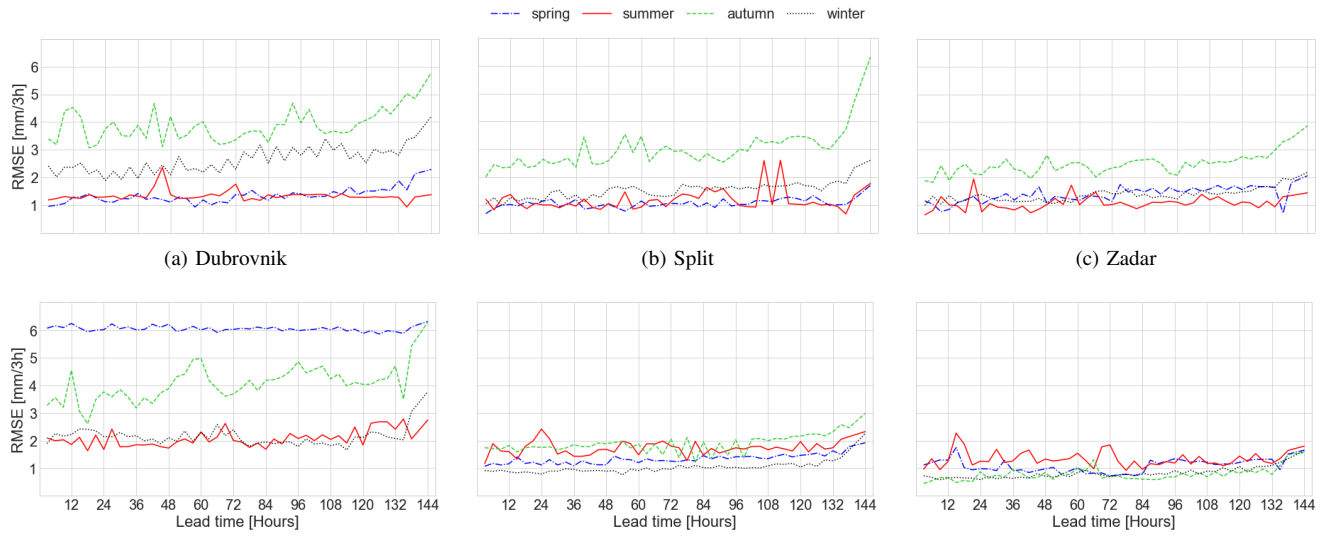
Considering seasonal differences, the improvement regarding temperature could be achieved for Split in the summer and Rijeka in the winter period (Figure 11). Spring and autumn precipitation forecasts could be improved for Rijeka, as well as for Split and Dubrovnik (Figure 12). Furthermore, the difference in time of the day is also evident (Figure 13). Improvement could be achieved for temperature forecasts, both



(a) Dubrovnik (b) Split (c) Zadar

(d) Rijeka (e) Zagreb (f) Osijek

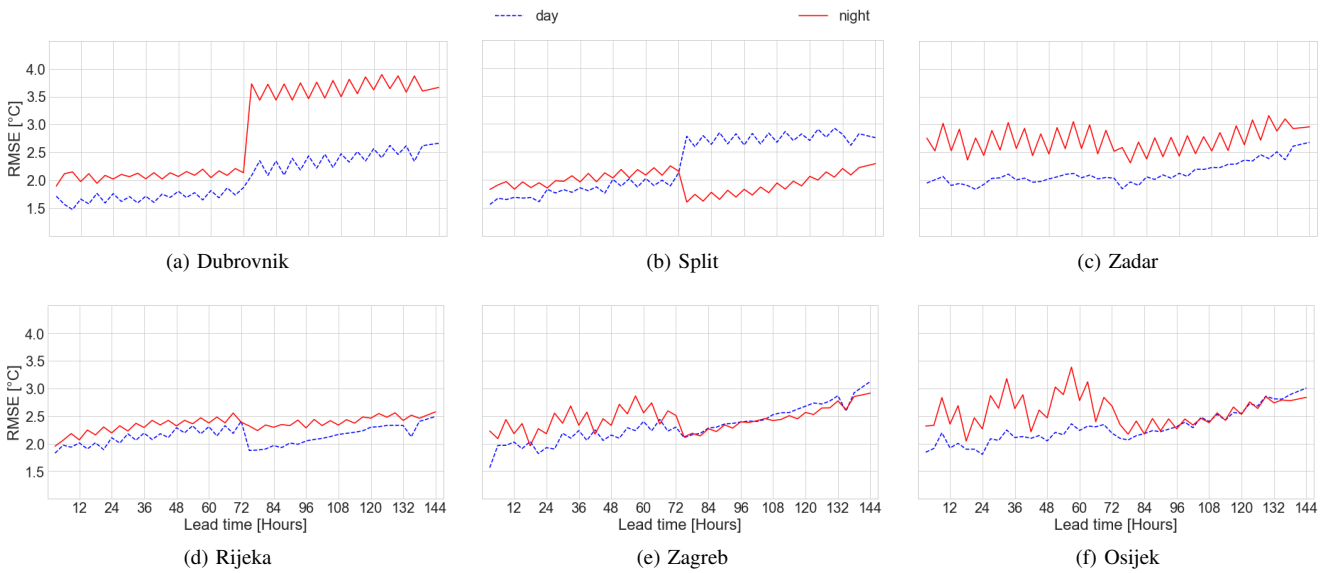
Fig. 11. Temperature error for different locations by seasons



(a) Dubrovnik (b) Split (c) Zadar

(d) Rijeka (e) Zagreb (f) Osijek

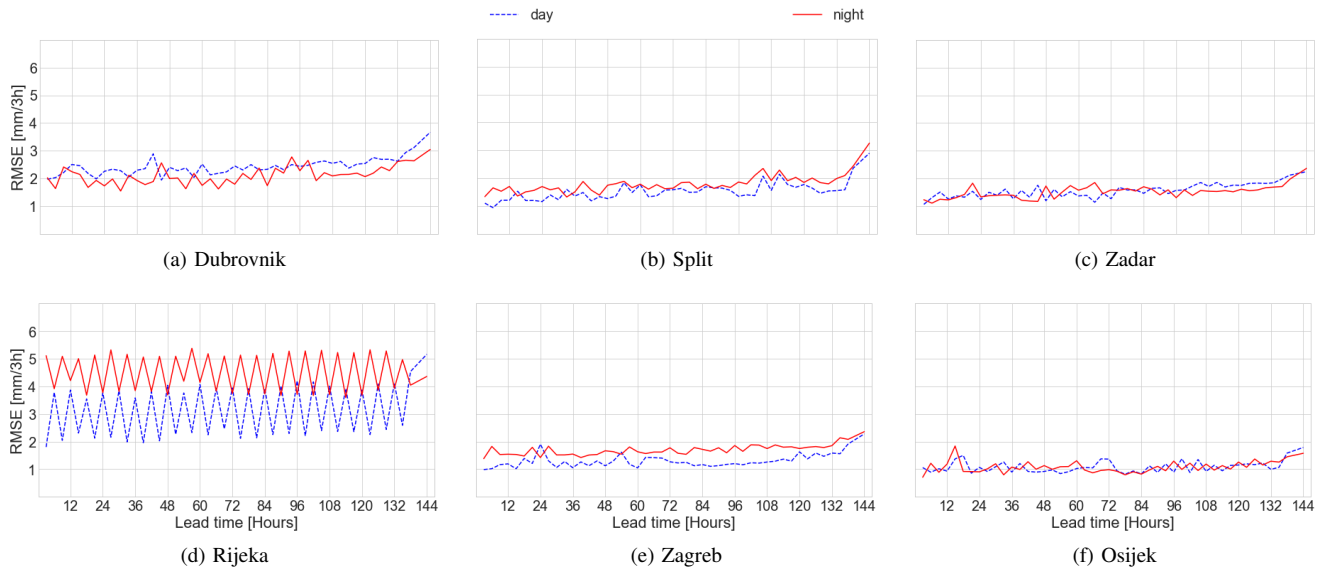
Fig. 12. Precipitation error for different locations by seasons



(a) Dubrovnik (b) Split (c) Zadar

(d) Rijeka (e) Zagreb (f) Osijek

Fig. 13. Temperature error for different locations by the time of a day



(a) Dubrovnik (b) Split (c) Zadar

(d) Rijeka (e) Zagreb (f) Osijek

Fig. 14. Precipitation error for different locations by the time of a day

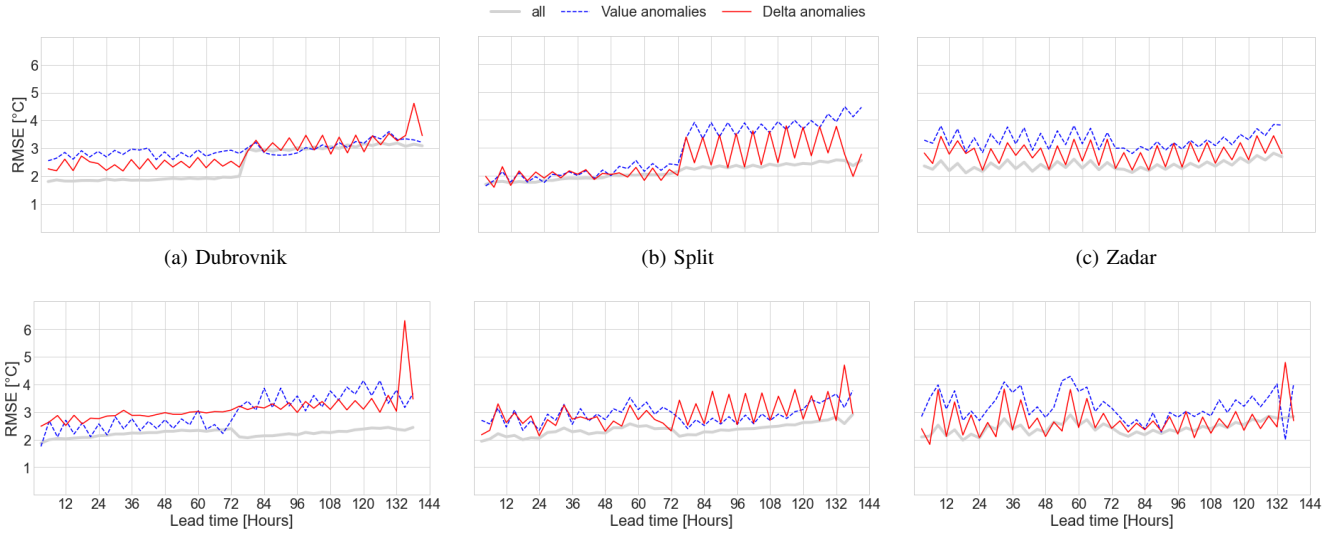
for Dubrovnik nightly forecasts and Split daily forecasts after the third day.

Finally, the evaluation of extremes shows that the extreme temperatures could be better predicted. This is especially visible considering the temperature in Split and Rijeka after the third day of forecast (Figure 15). Precipitation extremes could be predicted better for all the locations. This is especially evident for Rijeka (Figure 16).

To summarize, there is a potential for improvement for each location, time of the day, and season for both temperature and precipitation. Precipitation is generally the parameter that could be forecasted much better, especially considering locations with heavy precipitation that occurs often.

VII. Conclusion

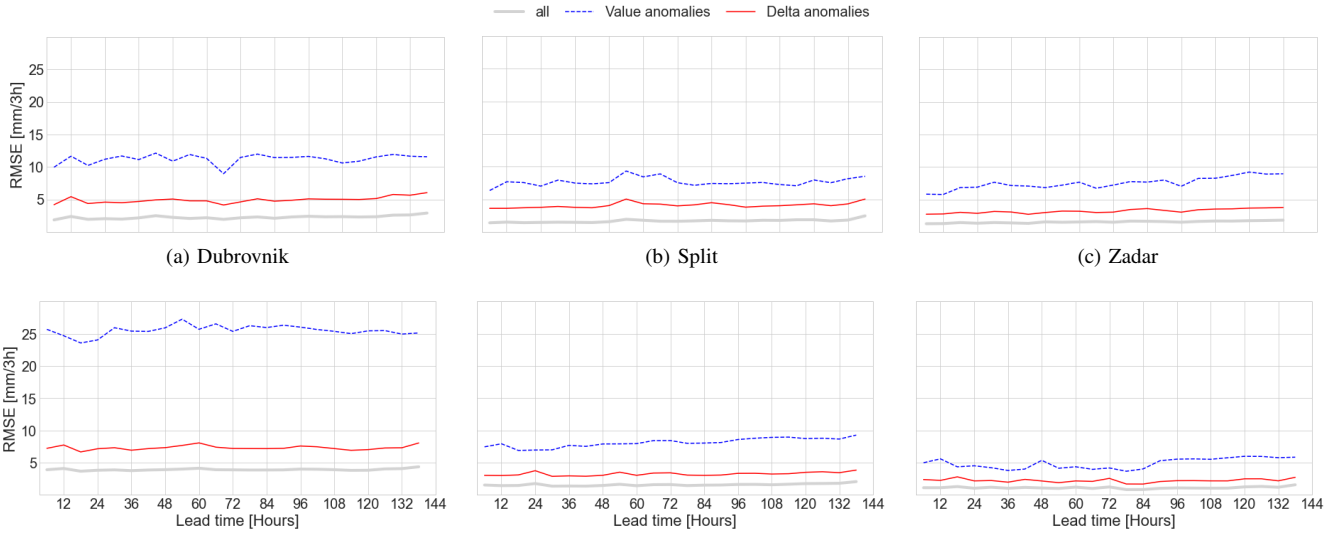
The paper has demonstrated the process of searching for scenarios in which the spatio-temporal model could be improved or replaced by a micromodel and deployed on Edge. The process was performed on the use case of a Numerical Weather Prediction model for temperature and precipitation. Black-box testing is used for evaluation, without taking into consideration the model structure. The dataset used for evaluation consists of outputs of temperature and precipitation for six locations in Croatia are measured and forecasted every three hours for 19 months. NWP was compared to the baseline models, namely model based on climatology and persistence model. The evaluation is performed considering the specificity of spatial and temporal components, in both normal and extreme conditions.



(a) Dubrovnik (b) Split (c) Zadar

(d) Rijeka (e) Zagreb (f) Osijek

Fig. 15. RMSE of NWP for extreme temperatures



(a) Dubrovnik (b) Split (c) Zadar

(d) Rijeka (e) Zagreb (f) Osijek

Fig. 16. RMSE of NWP for extreme precipitation

The results show that there is a room for the improvement in many scenarios. The detected scenarios can be the starting point for developing micromodels, which can be baseline models, machine learning models, etc. Pointing to the weaknesses of the existing model, such micromodels could result in better accuracy. Other benefits of using such models instead of the existing expert model are their speed and ease of deployment.

The process presented in this paper applies to any other spatio-temporal model and can mostly be performed regardless of the domain.

ACKNOWLEDGMENT

Authors would like to thank Ericsson Nikola Tesla Summer Camp students Iva Madunić, Veronika Ozretić and Marin Perić who performed the preliminary analysis.

REFERENCES

[1] A. Boyle, K. Aristovich, A. Adler -, C. Yang, X. Yan, Y. Wang, al, V. Melinda Gálfi, V. Lucarini, J. Wouters -, Z. Dong, and C. Guo, "A Literature Review of Spatio-temporal Data Analysis," *Journal of Physics: Conference Series*, vol. 1792, no. 1, p. 012056, feb 2021. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1792/1/012056https://iopscience.iop.org/article/10.1088/1742-6596/1792/1/012056/meta

[2] S. Shekhar, Y. Li, R. Y. Ali, E. Eftelioglu, X. Tang, and Z. Jiang, "Spatial and Spatiotemporal Data Mining," *Comprehensive Geographic Information Systems*, vol. 3, pp. 264–286, jan 2018.

[3] M. Schneider, "Spatial and Spatio-Temporal Data Models and Languages," *Encyclopedia of Database Systems*, pp. 2681–2685, 2009.

[4] W. Shi, G. Pallis, and Z. Xu, "Edge computing [scanning the issue]," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1474–1481, 2019.

[5] X.-B. Jin, R. Jonhson, R. Jeremiah, T.-L. Su, Y.-T. Bai, J.-L. Kong, X.-B. Jin, R. J. . Su, T.-L. . Bai, and Y.-T. . Kong, "The New Trend of State Estimation: From Model-Driven to Hybrid-Driven Methods," *Sensors 2021, Vol. 21, Page 2085*, vol. 21, no. 6, p. 2085, mar 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/6/2085/htmhttps://www.mdpi.com/1424-8220/21/6/2085

[6] F. J. Montáns, F. Chinesta, R. Gómez-Bombarelli, and J. N. Kutz, "Data-driven modeling and learning in science and engineering," *Comptes Rendus Mécanique*, vol. 347, no. 11, pp. 845–855, nov 2019.
[7] "The rise of data-driven modelling," *Nature Reviews Physics 2021 3:6*, vol. 3, no. 6, pp. 383–383, jun 2021. [Online]. Available: https://www.nature.com/articles/s42254-021-00336-z
[8] P. E. Ai, "A Generational Shift in AI PALANTIR EDGE AI OVERVIEW," Palantir Technologies Inc., Tech. Rep., 2022.
[9] P. Bauer, P. D. Dueben, T. Hoefler, T. Quintino, T. C. Schulthess, and N. P. Wedi, "The digital revolution of Earth-system science," *Nature Computational Science 2021 1:2*, vol. 1, no. 2, pp. 104–113, feb 2021. [Online]. Available: https://www.nature.com/articles/s43588-021-00023-0
[10] S. Nidhra, "Black Box and White Box Testing Techniques - A Literature Review," *International Journal of Embedded Systems and Applications*, vol. 2, no. 2, pp. 29–50, jun 2012.
[11] M. Gentry, "Micromodeling use cases and efficient implementation — insight," https://www.insight.com/en_US/content-and-resources/2021/micromodeling-use-cases-and-efficient-implementation.html, 2021, [Online; accessed 25-August-2022].
[12] A. Jindal, S. Qiao, R. Sen, and H. Patel, "Microlearner: A fine-grained Learning Optimizer for Big Data Workloads at Microsoft," *Proceedings - International Conference on Data Engineering*, vol. 2021-April, pp. 2423–2434, apr 2021. [Online]. Available: https://www.microsoft.com/en-us/research/publication/microlearner-a-fine-grained-learning-optimizer-for-big-data-workloads-at-microsoft/
[13] Lexalytics, "Machine Learning Micromodels: More Data is Not Always Better," https://www.lexalytics.com/blog/machine-learning-micromodels/, [Online; accessed 25-August-2022].
[14] E. Landau, "Introduction to micro-models or: how I learned to stop worrying and love overfitting," https://eric-landau.medium.com/introduction-to-micro-models-or-how-i-learned-to-stop-worrying-and-love-overfitting-fd8fbe98e99b, 2021, [Online; accessed 25-August-2022].
[15] S. Copei, C. Eickhoff, A. Malik, N. Nolte, U. Norbisrath, J. Sorgalla, J. Weber, and A. Zündorf, "From monolithic models to agile micromodels," in *Proceedings of the 10th International Conference on Model-Driven Engineering and Software Development - Volume 1: MODELSWARD,*, INSTICC. SciTePress, 2022, pp. 227–233.
[16] A. Lee, J. K. Kummerfeld, L. C. An, and R. Mihalcea, "Micromodels for Efficient, Explainable, and Reusable Systems: A Case Study on Mental Health," *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, pp. 4257–4272, sep 2021. [Online]. Available: https://arxiv.org/abs/2109.13770v1
[17] I. N. Kosovic, T. Mastelic, and D. Ivankovic, "Using Artificial Intelligence on environmental data from Internet of Things for estimating solar radiation: Comprehensive analysis," *Journal of Cleaner Production*, vol. 266, sep 2020. [Online]. Available: https://www.bib.irb.hr/1079839
[18] D. Džal, I. N. Kosović, T. Mastelić, D. Ivanković, T. Puljak, and S. Jozić, "Modelling Bathing Water Quality Using Official Monitoring Data," *Water 2021, Vol. 13, Page 3005*, vol. 13, no. 21, p. 3005, oct 2021. [Online]. Available: https://www.mdpi.com/2073-4441/13/21/3005/htmhttps://www.mdpi.com/2073-4441/13/21/3005
[19] T. Haiden, M. Janousek, F. Vitart, Z. Ben-Bouallegue, L. Ferranti, C. Prates, and D. Richardson, "Evaluation of ecmwf forecasts, including the 2021 upgrade," European Centre for Medium-Range Weather Forecasts (ECMWF), Tech. Rep., 2021.
[20] M. Tudor, S. Ivatek-Šahdan, A. Stanešić, Kristian Horvath, and A. Bajić, "Forecasting Weather in Croatia Using ALADIN Numerical Weather Prediction Model," *Climate Change and Regional/Local Responses*, may 2013. [Online]. Available: https://www.intechopen.com/chapters/42619
[21] C. Irrgang, N. Boers, M. Sonnewald, E. A. Barnes, C. Kadow, J. Staneva, and J. Saynisch-Wagner, "Towards neural Earth system modelling by integrating artificial intelligence in Earth system science," *Nature Machine Intelligence 2021 3:8*, vol. 3, no. 8, pp. 667–674, aug 2021. [Online]. Available: https://www.nature.com/articles/s42256-021-00374-3
[22] M. Bonavita, R. Arcucci, A. Carrassi, P. Dueben, A. J. Geer, B. Le Saux, N. Longépé, P. P. Mathieu, and L. Raynaud, "Machine Learning for Earth System Observation and Prediction," *Bulletin of the American Meteorological Society*, vol. 102, no. 4, pp. E710–E716, apr 2021. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/102/4/BAMS-D-20-0307.1.xml
[23] M. G. Schultz, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler, "Can deep learning beat numerical weather prediction?" *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, apr 2021. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0097
[24] C. Meteorological and H. Service, "Croatian meteorological and hydrological service webpage," https://meteo.hr, accessed: 2022-03-01.
[25] R. P. Ltd., "Reliable prognosis webpage," https://rp5.ru, accessed: 2022-03-01.
[26] E. Kalnay, "Atmospheric Modeling, Data Assimilation and Predictability," *Atmospheric Modeling, Data Assimilation and Predictability*, nov 2002.
[27] J. P. Gerrity, J. R. Gyakum, R. A. Anthes, L. F. Bosart, and E. A. O'Lenic, "Weather forecasting and prediction," *Access Science*, 2020. [Online]. Available: https://www.accessscience.com/content/weather-forecasting-and-prediction/742600
[28] M. Tudor, A. Stanešić, S. Ivatek-Šahdan, M. Hrastinski, I. Odak Plenković, K. Horvath, A. Bajić, and T. Kovačić, "Operational validation and verification of ALADIN forecast in Meteorological and Hydrological Service of Croatia," *Croatian Meteorological Journal*, vol. 50, pp. 47–70, 2015. [Online]. Available: https://hrcak.srce.hr/155403
[29] R. G. Owens and T. Hewson, "Ecmwf forecast user guide," European Centre for Medium-Range Weather Forecasts (ECMWF), Tech. Rep., 2018. [Online]. Available: https://www.ecmwf.int/en/elibrary/16559-ecmwf-forecast-user-guide
[30] I. Odak Plenković, I. Schicker, M. Dabernig, K. Horvath, and E. Keresturi, "Analog-based post-processing of the ALADIN-LAEF ensemble predictions in complex terrain," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, pp. 1842–1860, 2020. [Online]. Available: https://doi.org/10.1002/qj.3769
[31] E. C. for Medium-Range Weather Forecasts, "European centre for medium-range weather forecasts webpage," https://www.ecmwf.int/en/about, accessed: 2022-06-06.
[32] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, "WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 11, p. e2020MS002203, nov 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1029/2020MS002203https://onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002203https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020MS002203
[33] ECMWF, *IFS Documentation CY43R3*, ser. IFS Documentation. ECMWF, 2017.
[34] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.
[35] W. M. Organization, "Climate explorer webpage," https://climexp.knmi.nl, accessed: 2022-03-01.

Ivana Nižetić Kosović is a researcher at Ericsson Nikola Tesla d.d., Research Department. She obtained her diploma in mathematics at the Faculty of Science (University of Zagreb, Croatia) and completed her PhD at the Faculty of Electrical Engineering and Computing (University of Zagreb, Croatia), where she was working as an Assistant Professor. Her scientific interests include spatio-temporal reasoning, artificial intelligence and heterogeneous data analysis.

Toni Mastelić is a researcher at Ericsson Nikola Tesla d.d., Research Department. He did his bachelor and masters studies in Computer Science at the University of Split, FESB, Croatia, where he received his Bachelor's degree in 2009, and Master degree in 2011. Afterwards, he worked as a researcher and later on as a University Assistant at Vienna University of Technology, Austria, where he pursued his PhD. Finally, he received his PhD degree in 2015 at the Institute of Software Technology and Interactive Systems, Vienna University of Technology.

Domina Sokol is a student at the University of Split. She received her Bachelor's degree in Mathematics and Computer Science at the Faculty of Science, where she is currently pursuing a Master's degree in Mathematics. Her fields of interest include Artificial Intelligence, Computational Linguistics, and Data Science.

Diana Škurić Kuraži is a researcher in the Research Department of the Ericsson Nikola Tesla company in Croatia. She received a Master's degree in Physics and Geophysics from the Department of Geophysics, University of Zagreb, Croatia in 2015. She is PhD candidate in the same Department of Geophysics with the dissertation topic "Enhanced method of forest fire risk assessment" and a member of Croatian Meteorological Society and Croatian Agrometeorological Society. Her research interests are forest fires, forest fire warning systems, numerical weather prognostic models and data analysis.