Consortium

for

Small-Scale Modelling

Technical Report No. 50

The COSMO Priority Project AWARE: Appraisal of "Challenging WeAther" FoREcasts Final Report

November 2023

DOI: 10.5676/DWD_pub/nwv/cosmo-tr_50

www.cosmo-model.org

Editor: Massimo Milelli, CIMA Foundation

The COSMO Priority Project AWARE: Appraisal of "Challenging WeAther" FoREcasts Final Report

Flora Gofa[1*], ***Anastasia Bundel***[2*], ***Maria Stefania Tesini***[3], ***Chiara Marsigli***[4], ***Michael Hoff***[4], ***Dimitra Boucouvala***[1] ***Andrzej Mazur***[5], ***Joanna Linkowska***[5], ***Grzegorz Duniec***[5] ***Daniel Cattani***[6], ***Benoit Pasquier***[6], ***Anatoly Muraviev***[2] ***Ekaterina Tatarinovich***[2], ***Yulia Khlestova***[2], ***Denis Zakharchenko***[2]

[*] ***Project Coordinators***

[1] ***HNMS***

[2] ***RHM***[‡]

[3] ***ARPAE-SIMC***

[4] ***DWD***

[5] ***IMGW-PIB***

[6] ***MeteoSwiss***

[‡] Contribution received before February 24, 2022

1 Introduction

The increased demand to provide accurate forecasts of extreme weather leads to the question to how objectively evaluate such forecasts. The main weather parameters of interest are: thunderstorms (heavy precipitation, lightning), severe wind (and wind gusts), min-max temperature (persistence), visibility (fog), extreme convective phenomena like tornadoes, dust-devils, clear-air turbulence (CAT) etc. In COSMO consortium, there have been several studies partially related to challenging weather (CW) aspects. However, up to now there haven't been a project explicitly focusing on evaluation and development of HIW forecasts.

As the resolution of state-of-the-art NWP models is growing, there is more detailed and precise information on the variables necessary for calculating, for example, the electrical properties of the atmosphere (temperature, humidity, wind, ice, water content of particles, etc.). The physical and microphysical processes leading to local CW of convective nature are better reproduced in modern NWP models. This should improve the direct forecasting of CW of convective nature, such as thunderstorms, hail, squalls, and showers. However, the models cannot satisfy yet all the needs for CW predictions. Thus, different postprocessing methods are required. However, it is important to compare the forecasts based on direct model output (DMO) and postprocessing, where it is feasible.

Despite significant progress in short-range forecasting, HIW continues to cause most part of damage to the economy and society, up to losses of human lives. WMO initiated a project dedicated to HIW research (*WMO HIW implementation plan*). The overall objective of the High Impact Weather (HIWeather) project is to: "Promote cooperative international research to achieve a dramatic increase in resilience to high impact weather, worldwide, through improving forecasts for timescales of minutes to two weeks and enhancing their communication and utility in social, economic and environmental applications". One component of the project is the verification and the research is focused on approaches with relevance to hazard predictions. **PP AWARE was part of the WMO HIW project, and a short report of its main advances was included in HIWeather quarterly newsletter.**

Forecast methods and verification are important aspects of any CW consideration. While traditional verification methods have limited usefulness in this context, many of the newer diagnostic approaches may provide useful information to aid understanding of errors in model processes for such weather regimes. Verification of multi-scale prediction of CW has much in common with routine verification performed at most national meteorological centres. It mainly concerns surface variables such as precipitation, wind, temperature, etc., using both point-wise and spatial approaches to meet the needs of a variety of users. On the other hand, CW phenomena as fog or lightning are usually not directly forecasted by NWP models, and thus appropriate empirical methods are applied for their prediction.

Known deficiencies to be addressed are:

1. Models may not capture the intensity of high impact events (sub-grid scale processes, coarse resolution, difficulty representing processes)
2. Often a mismatch between what models can provide and what warnings need to be made for: Lightning, hail, wind gusts, fog, etc.
3. Large uncertainty with extreme events (Ensemble/probabilistic forecasts to measure "extremeness").

Several new verification methods have been proposed for evaluating the spatial structures simulated by high resolution models and this remains an active area of research. While most of these spatial methods measure forecast quality, some of them (e.g., variograms) address the realism of the forecast, which may be of particular interest to modellers. Spatial verification approaches are now starting to be applied to high resolution ensemble forecasts, but much remains to be done to understand what can be learned from these approaches, both in terms of quantifying ensemble performance, and in calibrating and postprocessing ensembles to improve forecast quality and utility. The utility of spatial verification for evaluating hazard impact forecasts (e.g., flood inundation, fire spread, blizzard extent and intensity, pollution cloud) needs to be explored, especially since graphical advice and warnings are becoming more common.

The goal of the PP-AWARE was to provide COSMO Community with an overview of forecast methods and forecast evaluation approaches that are linked to high impact weather (not necessarily considered extreme to all users). As an outcome of this project, the whole chain of observing and predicting CW/HIW, as well as evaluating, and distributing CW/HIW forecasts was studied; a number of most successful and promising methods was identified and developed based on the experience of the COSMO countries and the study of the state of the art in the world.

The importance of accurate forecasting of challenging weather occurrences is obvious. With the term challenging weather (CW) or high impact weather (HIW), we consider the events the local society is not routinely accustomed to experiencing. Such events could be extreme in amplitude (intense winds, or heavy convective precipitation), rare (lie in a tail of climatological distribution for a particular location) or high impact by being prolonged 'regimes' (droughts, heatwaves or cold spells), while others even if not very rare can be considered challenging if society is particularly vulnerable to them (e.g. impact of fog on transportation). In theory, a weather event could be high-impact when it is inherently less predictable and society does not have sufficient forewarning to take mitigating action.

Key forecast quality and verification aspects that were considered in this project include:

1. How well high-impact weather is represented in the observations, including biases and random errors, and their sensitivity to observation density.
2. How well high-impact weather is represented in models, including systematic and stochastic errors, and their sensitivity to model resolution.
3. How well high-impact weather is represented in postprocessing.
4. The predictability, current predictive skill, and the user's interpretation of forecast value in high-impact weather situations.

2 The Main Outcomes of AWARE Project

The project was focused on several HIW parameters, basically, events of convective origin (intense precipitation and flooding related to it, lightning, tornadoes), and to some extent on the visibility prediction. It was planned initially to broaden the scope of the project to include wind gusts, clear-air turbulence, possibly dust forecast, etc. The lack of resources and unpredictable events, such as COVID-19 pandemic, somewhat reduced the project field. Nevertheless, the outcomes of the project provided a substantial basis for further research. In the following chapters, the reports that were prepared from the various participants as deliverables to the Task work, are listed. The summary of the work concluded in PP-AWARE can be given in the following remarks:

Task 1 (Chapter 4) is basic for the understanding of the nature of phenomena studied within the project. The task considers which observations are necessary to verify HIW forecasts, as well as issues related to observation sparseness, quality, and thresholds. Furthermore, through the Task, some work effort was given to identify observation requirements for monitoring of selected hazards and for assessing forecast accuracy and quantifying the role of observation uncertainty. The study on observation uncertainty was initiated within the INSPECT project (COSMO technical Report 37, Chapter 4.1.3 and Chapter 5). The outcome of this task is the description of available HIW observations (including non-standard ones) and their characteristics.

Task 2 (Chapter 5) discusses that the verification of many HIW events requires metrics that remain useful for rare events. Their main characteristics are that the metrics must be less dependent on the climatology of the event. The dependency on spatial and temporal scales and sampling of observational data should be minimized, and the dependency on the verification grid should be minimized as well. Hits and false alarms should be taken into account. As no single score exists that addresses all these properties, the response of commonly used scores on HIW for these properties was studied on selected test cases. Scores behaviour for the evaluation of both the deterministic and ensemble forecasts of HIW (SEDI, EDS, EDI, SEEPS, CRPS) provided a more fair approach for the evaluation of high resolution precipitation events as it took into account the variable climatology of the model domain. An important part of this Task was the application of Extreme Value Theory resulted in a rather sophisticated approach to evaluate contiguous precipitation areas observed and predicted by the radar-based precipitation nowcasting system. The Generalized Pareto distribution (GPD) was chosen to fit precipitation object area maxima distribution. We introduced a new measure of the forecast quality based on the intersection of GPD parameters in the nowcasting results and in the observations. Based on this measure, we made some conclusions about the nowcasting quality in Central Russia. As extracted from the analysis, the EVT is applicable to such objects only with a clear understanding of the theoretical prerequisites and using suitable statistical methods and reliable data processing tools. Otherwise, the results obtained may be useless, accidental, or even harmful.

The Task 3 (Chapter 6) makes use of the analysis and outcomes of the previous Tasks and is also connected with and continued from PP-INSPECT and MesoVICT project on the spatial methods. Neighbourhood methods continue to be the basic spatial approach in operational practice. Different comparisons of neighbourhood and object-based scores as applied to HIW variables of convective origin (LPI, intense precipitation, reflectivity) are examined. These two groups of spatial methods provide a comprehensive framework for the operational verification in the forecast centres. MODE was found the most promising object-based approach to the evaluation of ensemble forecast of convective events (Chapters 6.3 and 6.6). MODE is flexible and tunable, it can be run with matching forecast and

observations object, but it doesn't require matching necessarily. It is applicable to both deterministic and ensemble fields.

With respect to the Lightning Potential Index, it played a key role in the analysis of convective events (Chapters 6.2 and 6.5). However, we noted the significance to derive upscaled LPI products in order to gain reliability in the forecasts. LPI raw values need to be appropriately filtered and thresholded according to the area and period or season examined or even type of the weather event. The filters in LPI formulas are more efficient in some models compared to others, but they show inefficient pruning of spurious signal especially in orographic precipitation. The LPI produced reliable lightning information during stronger storms, much like observed in observational data while it was shown that compared to thermodynamical indices forecasts, LPI is somewhat better at distinguishing lightning-producing storms and this may be of importance to some user groups.

With respect to ensemble forecasts of extreme precipitation, one of the proposed methodology was the DIST spatial method (Chapter 6.4). It permits the use of gridded high-resolution rain-gauges network values, but gridded observations, such as radar precipitation analysis, can be used as well. The main advantage of this approach is that no precipitation analysis is required and information about localized maxima of precipitation can be considered, as well as the variability of the precipitation field inside the area of interest.

Finally, the results of SINFONY project are shared within PP-AWARE (Chapter 6.6), a project that is dealing with the development, adaptation, and operationalization of innovative, spatially based verification methods of the entire process chain of the integrated forecasting system consisting of data assimilation, nowcasting and numerical short-term prediction. In the running PP-AWARE period, we tested various verification metrics, existing as well as new ones, such as verification metrics based on MMI (pseudomember by Johnson et al. (2020). When using a 40 member object ensemble from NWP, nowcasting and combined products, the number of existing objects could become massively huge and not manageable without applying filter methods like pseudomembers.

Task 4 (Chapter 7) is devoted to postprocessing for HIW, such as intense precipitation for flooding prediction, fogs, flash rate, and tornadoes risk. Taking into account the development of atmospheric modelling, the forecast of the horizontal visibility in the postprocessing of the model output (including machine learning methods) seems to be the most appropriate option. It is also physically justified, since it is based on the prognostic cloud characteristics and/or parameters of environment. We can apply a set of parametrizations of horizontal visibility thus creating a kind of ensemble. The fog is a meteorological phenomenon with a high degree of locality. We try to reduce the prognostic error by using a set of postprocessing approaches.

One of the most dangerous HIW event is tornado. Chapter 7.1 describes the experience in postprocessing COSMO results for predicting areas with tornado risk. This experience can be ported to ICON model output.

Some tools have been developed to provide mean, maximum and some other percentile values of the precipitation field over the catchment areas of the Emilia-Romagna region. Exceeding predefined thresholds can give useful indications for situations of intense precipitation possibly leading to floods. Probabilistic products help forecaster to assess confidence in one modeling chain or the other. Deterministic products for each warning area are validated on a seasonal basis using "bubbles plot" charts, a sort of the scatter plot in which the data points are replaced with bubbles and the sizes of the bubbles are determined by the number of events. The advantage of this approach is that the nature of the forecast errors can more easily be diagnosed.

3 Future Work

PP-AWARE was a particularly extensive project that dealt with a broad range of issues related to high impact weather. Consequently, many parts were analysed only partially or not at all when there was no active contribution due to lack of human resources. It provided however a better understanding of the areas that is worth investing more effort to approach through a next project or a dedicated WG5 research activity. It is worth mentioning that in the guidelines that define the future work of WG5, there are actions that are connected to the knowledge gained through AWARE.

The following list summarizes the work that could define the next phase of a project that deals with the evaluation of such weather:

HIW phenomena: visibility range (fog), discrimination between severe and non-severe convection, extreme temperatures and winds,

Observation Types: application of non-conventional observations

Methodologies: Multivariate verification statistics (several gridpoints-leadtimes-variables in all possible combinations with respective το obs), further study on obs uncertainty with application to scores, Impact-based warnings issuing and evaluation

Models: application on convection permitting ensemble systems.

More specifically, there are two main directions of future work:

1. Stressing of observations role in HIW through the use of non-conventional observation types in the evaluation of HIW phenomena and the impact of observation uncertainty on the statistical scores.
2. Verification scheme for convection permitting ensemble forecasts that could include the building of a robust verification framework that is based on object-based approaches.

4 Challenges in Observing High Impact Weather (HIW)

Question: How well high-impact weather is represented in the observations, including biases and random errors, and their sensitivity to observation density?

HIW phenomena studied: visibility range (fog), thunderstorms (w. lightning), intense precipitation, extreme temperatures and winds.

4.1 Overview of Challenging/High Impact Weather observational data sources characteristics - Review of non conventional observations and their use in verification

Andrzej Mazur, Institute of Meteorology and Water Management – National Research Institute

Chiara Marsigli, DWD

Anastasia Bundel, RHM

This task considers which observations are necessary to verify HIW forecasts, as well as issues related to observation sparseness, quality, and thresholds. HIW prediction improvement depends crucially on availability of dense observations. The uncertainty is higher in new types of observations, and it becomes necessary to take it into account. The overview of methods to account for observation uncertainty is considered in paragraph 1.2. Often, the best way is to use several observational datasets to this purpose. For verification and postprocessing, the essential step is to find good correspondence between the forecast and observation, or reference. In [C. Marsigli et al, 2021], a framework for the verification of high-impact weather is proposed, including the definition of forecast and observations in this context and creation of a verification set. This was discussed at the IVMW2020 [https://jwgfvr.univie.ac.at/]. It was noted by T.Bullock [https://www.univie.ac.at/img-wien/jwgfvr/2020IVMWO_Outcomes&Photo Mosaic.pdf] that there is always some processing (both on observations and forecast) for enabling comparison. We need just to be clear on what is being done to the model output and/or obs prior comparison (e.g., conversion of radar reflectivities to rainfall rate, versus forward model to reproduce radar reflectivities).

It can be said that every weather has its impact. Starting with the least inconvenient, like:

1. inconvenience of carrying an umbrella/sun glasses,
2. higher power bills,

through moderately troublesome:

1. possibility of dispersion of atmospheric pollutants,
2. flight delays due to weather conditions

to very dangerous in consequences, like:

1. catastrophes in sea, land and air traffic
2. destruction caused by a flood or a tornado

To someone affected, any of these may seem "significant" at that moment. Some impacts are clearly more significant than others. There are four general categories of impacts:

1. Low-impact – minor inconvenience, small and local economic losses, etc.
2. Moderate-impact – minor damage, some social disruption, etc.
3. High-impact – damage, risks to health, broad economic impact, etc.
4. Extreme-impact – dramatic losses, deaths, injuries, major social disruption, etc.

Since every (kind of) weather has its impact, each weather element can be treated as an impact source. It's just a matter of scale.

1. "regular" elements – temperature, precipitation, wind speed...
2. "specific" elements – visibility limitations, thunderstorms, tornadoes, ...

Observational data for each element can be obtained from a variety of sources. The main sources can be divided into:

1. ***Data from SYNOP stations***
2. ***Lightning Detection Networks (LDN)***
3. ***Radar data, Doppler radar data***
4. ***Satellite products***
5. ***Nowcasting products used as reference data***
6. ***Non-conventional data such as datasets derived from telecommunication systems, data collected from citizens, reports of impacts and claim/damage reports from insurance companies, social networks, data from cameras and images***
7. ***Other data***

Below, an overview of these sources is given. This overview is far from being exhaustive, and, according to the purposes of PP AWARE, is focused on the types of observations used in the project tasks, namely, events of convective origin (extreme precipitation, lightning, convective cells, tornadoes) and fog. In [C. Marsigli et al, 2021], an overview of new observation types is given in more detail.

1. (a) **Data from SYNOP stations***, climatological stations, rain gauges, telemetry stations includes measurements of, among others, the following values: *temperature, precipitation, visibility range/limitations, wind speed, wind gusts, occurrence of fog/haze, occurrence of thunderstorm with lightning (limited to a remark as "day with lighting" or similar).*

* *An exemplary information from European/Polish SYNOP station after decoding a SYNOP (encoded) wire*
rrrr mm dz gg number n dd ff vv ww w1w2 pppp ttt nh cl h cm ch tdtdtd a ppp rrr
2020 3 3 6 01001 7 120 6 10 2 22 1013.2 1.1 7 5 3 −1000 −1000 −3.7 7 −0.6 0
tntntn txtxtx tgtg sss ff_911 ddd ss_931 statists ff_910 p0 rrr_24 −0.2 −1000 −1000 0 12 −1000 −1000 −1000 −1000 1012 0

These conventional observations remain the basic source of data for many HIW events, e.g., extreme precipitation, extreme temperatures and wind. They pass thorough quality control and are regular in time. There are long time series of synoptic measurements, which is important in the study of rare phenomena. However, the problem with these stations (both manned and unmanned) is that the measurement is valid only for the location of a particular station. The representativeness may be (artificially) extended up to some dozens of kilometres, but it is not necessarily valid for example for stations located in complex terrain etc. Some specific measurements (like fog/visibility range[†]) are being transferred, however, to more universal, mobile installations. Data of SYNOP stations: visual thunderstorm occurrence at a given obs time and between obs times in a radius of 5 km.

Another problem with SYNOP observations is that they often do not permit full characterization of specific HIW phenomena, such as visibility limitations, thunderstorms, tornadoes. Thus, in Europe, 10 years ago, a list of new weather elements to be subject to routine verification was proposed by [Wilson and Mittermaier 2009]. Among others, visibility/fog, atmospheric stability indices and freezing rain were mentioned, and the observations needed for the verification of these additional forecast products were reviewed.

A general remark regarding LDN, and (even more) especially radar or satellite data, is as follows: for their correct use, a proper software is needed that will allow the data to be transferred to the appropriate (required) format.

1. (a) **Lightning Detection Networks**

Lightning Detection Networks (LDNs) are based on lightning detectors that indicate electrical activity. The basic assumption made when creating LDN ensures that due to proper triangulation, it is possible to estimate the almost exact location of the flash. LDNs can detect dry thunderstorms. Furthermore, lightning detectors do not suffer from a masking effect and provide confirmation when a shower cloud has evolved into a thunderstorm.

If used as a proxy for a thunderstorm, a question arises: How many strokes are needed to detect the occurrence of a thunderstorm? The matching of the two entities in the verified pair should be checked before the computation of summary measures. Any thresholds used to identify the objects of the two quantities must also be studied to ensure that the identification and comparison is as unbiased (from the observation point of view) as possible [C. Marsigli et al., 2021]. In the present report, verification using LPI (lightning potential index) and LDN data is studied in Tasks 3.1 and 3.2 (Chapter 6).

Global LDN: websites

The most popular global resources about lightning are:

1. https://blitzortung.org, a worldwide social network for determining location of lightnings in real time. In figures below exemplary screenshots from the webpage in static and dynamic presentation.

[†]Haltere, Nicolas et al., 2006, Automatic fog detection and estimation of visibility distance through use of an onboard camera, Mach. Vis. Appl., 17, 8-20, 10.1007/s00138-005-0011-1

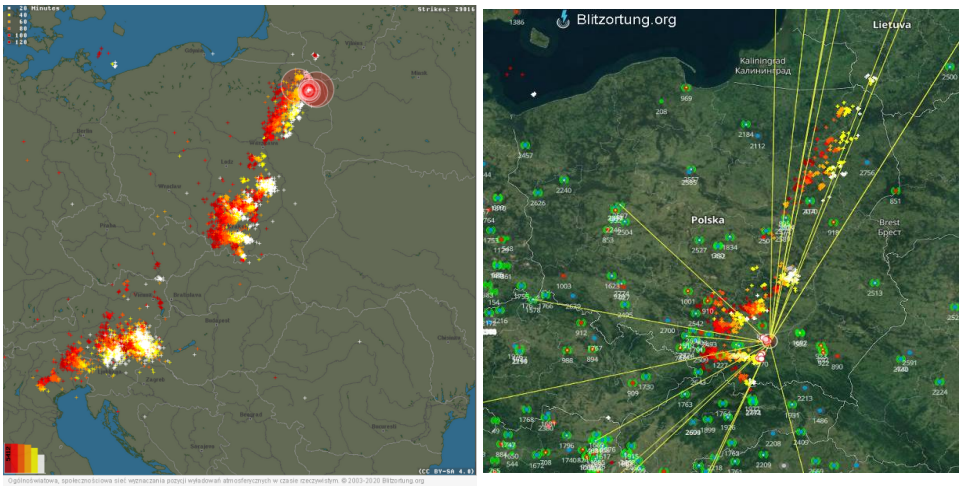


Figure 1: Webpage https://blitzortung.org. On the left – standard discharge image – locations marked with crosses, the more red the crosses are – the older occurrence of lightning. On the right, a dynamic map with additionally marked locations of the detectors and lines to the detectors that detected a specific discharge.

1. http://wwlln.net/TOGA_network_global_maps.htm: Very Low Frequency sensors. Lightning stroke positions are shown as colored dots which "cool down" from blue for the most recent (occurring within the last 10 min) through green and yellow to red for the oldest (30-40 minutes earlier).

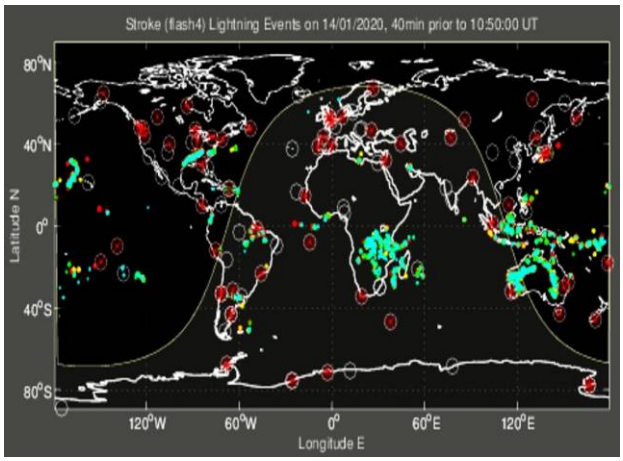


Figure 2: http://wwlln.net/TOGA_network_global_maps.htm

Regional LDNs

Regional lightning detection networks: Very Low Frequency sensors in the real time within 100-300 km radius, detect two types of lightnings: cloud-earth and intra-cloud. In Poland LDN operated by NWS is called PERUN. It is basically identical to French SAFIR. (Surveillance et Alerte Foudre par Interférometrie Radioélectrique).

An exemplary information from PERUN LDN: time, location, flash type, intensity etc.

11/01/2011 00:24:50;0004FCFFFFFFFFFFFFFFFF;1;0;7561;538442;193454;0;218;0;0;0;0;0;0;7;10;0

11/01/2011 00:25:58;0004FCFFFFFFFFFFFFFFFF;1;0;6839;537325;196241;0;218;0;0;0;0;0;0;7;10;0

11/01/2011 00:26:35;0004FCFFFFFFFFFFFFFFFF;2;0;8280;536018;194977;0;203;0;0;0;0;0;0;7;10;0

11/01/2011 00:26:35;0004FCFFFFFFFFFFFFFFFF;2;0;8788;536502;190226;0;103;0;0;0;0;0;0;5;2;0

In Russia, the lightning detection system of Roshydromet ALVES 9.07 is used [Gubenko I. 2016; Snegurov A.V., Snegurov V.S. 2012]. In [Gubenko I. 2016], it is shown that the accuracy of regional Russian LDN is higher than WWLLN (comparison to SYNOP data).

Other examples of European LDNs are BLIDS (which stands for Blitz-Informationsdienst von Siemens), FLITS (in Netherlands and Belgium) or LINET, developed in Munich, Germany. In [C.Marsigli et al. 2021], other lightning detection networks are listed, and references to works with applications of LDN data in verification are given including spatial approach and combining different data sources.

1. (a) **Radar data, Doppler radar data**

-Precipitation intensity and type, wind speed, lightning

Radar data and/or Doppler radar data are acquired from weather radar that indicates precipitation (in a standard mode) and wind field (in Doppler mode). Both phenomena are associated with thunderstorms and can help indicate storm strength. In general, weather radar will show a developing storm before a lightning detector does. However, weather radar also suffers from a masking effect by attenuation, where precipitation close to the radar can hide precipitation farther away. Moreover, if there is no precipitation (at all), availability of radar data declines rapidly in both standard and Doppler mode. This situation may occur in connection with the phenomenon of so-called dry thunderstorm. In this case lightning(s) may be also located outside any precipitation recorded by radar.

In addition to stationary (ground-located) installations for the detection of flashes, mobile devices are also used and carried on ships or airplanes. Large airliners are more likely to use weather radar than lightning detectors, since weather radar can detect smaller storms that also cause turbulence. Modern avionics for additional safety include lightning detection as well. For smaller aircraft, especially in general aviation (where the aircraft nose is not big enough to install a radome) lightning detectors can find and display IC and CG[‡] flashes.

Digital radar systems now offer thunderstorm tracking surveillance. This provides users with the ability to acquire detailed information of each storm cloud being tracked. Thunderstorms are first identified by matching precipitation raw data received from the radar pulse to some sort of template preprogrammed into the system. In order for a thunderstorm to be identified, it has to meet strict definitions of intensity and shape that distinguish it from any non-convective cloud. Usually, it must show signs of organization in the horizontal and continuity in the vertical: a core (more intense center) to be identified/tracked by digital radar trackers.

Radar reflectivity fields are used for the estimation of the risk of tornadoes, and for verification of these events (see Task 4.1.2).

1. (a) **Satellite products**

Occurrence of fog/haze, detection of convective storms, cloud properties (direct measurement of moisture and instability[§]), also via convective indices and CAPE

[‡] $IC --$ *inter – cloud lightning,* $CG --$ *cloud – to – ground flash*

[§] *infrared* (IR) 10.8 μm *and water vapor* (WV) 6.2 μm *channels*

An advantage of the satellite products is that they provide data over data-sparse regions. Satellite data detection of convective storms is based on direct measurement of moisture and instability:

Intensity = IR + ((IR-NWP)-(WV-IR))[¶]

with IR, NWP, WV being temperature obtained from different channels.

From the above equation, it is necessary to use the PA (e.g., the results of the global GFS model).

Convective indices: in general, can be a good prognostic tool if only forecasters could understand why values are approaching critical levels, like in the examples below:

1. (a) i. A. Showalter Index – extreme instabilities for SI less than -6
   B. Total Totals Index – severe storms with TTI greater than 50
   C. K Index – high convective potential for K greater than 40
   D. SWEAT Index – severe phenomena possible for SWEAT greater than 300
   E. Lifted Index – extreme instabilities for LI less than -6
   F. CAPE – extreme values of 2500 and more

An example of thunderstorm verification for clouds based on satellite data is given in [Keller et al. 2015].

In RHM, a study on identification of the areas of deep convection based on satellite data is carried out [Shishov A.E., I.A. Gorlach 2020; Shishov A.E. 2021]. Based on calibrated radiative temperature from Seviri, Meteosat-11, using a threshold, a mask of deep convection areas is found. Then the cell shape is determined. The cells are traced in time based on the normalized overlapping area. Cell destroying is also taken into account. Then, the cell movement direction, deformation, and other characteristics are identified. Figure 3 gives an example of the areas of deep convection in the visualization system developed by the authors. It is planned to involve other data for deep convection area identification, such as surface obs (KH01, METAR) and COSMO-Ru / ICON-Ru prognostic fields. It is planned to study the feasibility of using this product as a reference for verification of a model analogue.

[¶]da Silva et al., 2016. A method for convective storm detection using satellite data. Atmòsfera, 29 (4), 343-358

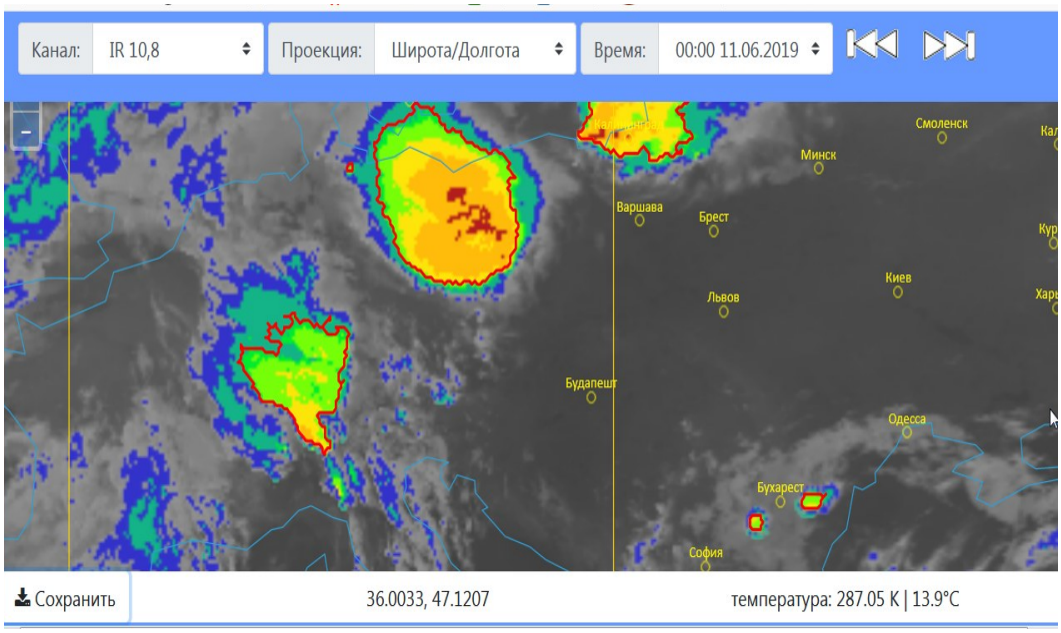


Figure 3: An example of the areas of deep convection in the visualization system.

Satellite products are now widely used to derive the information about the **fogs and low stratus**, besides SYNOP reports containing visibility range/limitations. Problems of visibility measures from manual and automatic stations are described in [Wilson and Mittermaier 2009]. The main problem of point observations is that they are scarce and not sufficient to reproduce the spatial structure of fog.

In [Morales at al. 2013], verification is performed for low clouds in the model as proxy for fog *vs* cloud type product from satellite NWC-SAF as observations. In [Ehrler 2018, Westerhuis et al. 2018], liquid water path (LWP) in the model is compared vs satellite data (channel combination) to give a Cloud Confidence level. A paper is under preparation by the Russian team (N. Chubarova, Yu. Khlestova, et. al.), which compares model LWP using one- and two-moment physics COSMO scheme with satellite product.

Satellite images also enable reconstruction of tornadoes tracks by fallen trees (see also Task 4.1.2).

1. (a) **Nowcasting products used as reference data**

National Meteorological Services develop tools for nowcasting, where data from different sources (satellite, radar, lightning, etc.) are integrated in a coherent framework. The detected variables/objects of nowcasting (such as thunderstorm cells, hail) can become observations against which to verify the model forecast. Thus, nowcasting products are proposed as observed data instead of prediction tools if we consider step 0 of the nowcasting algorithm as an "analysis".

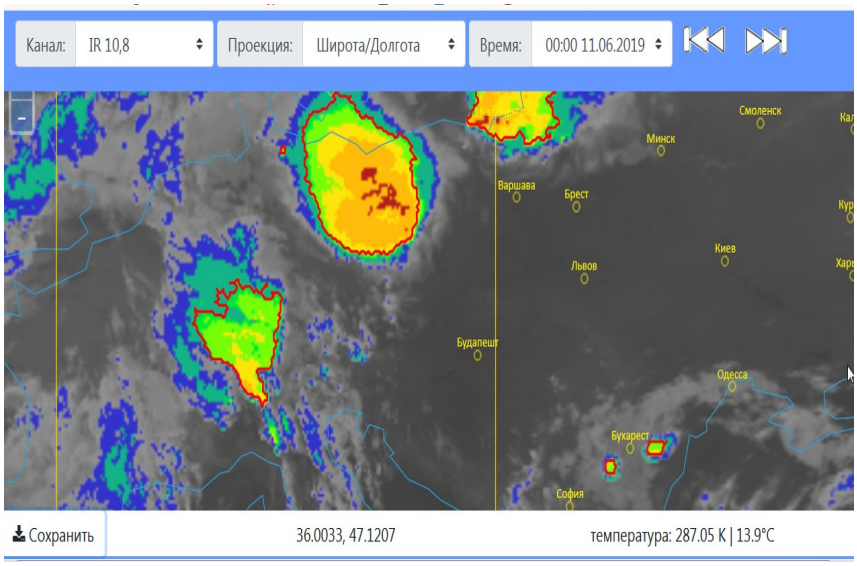


Figure 4: Nowcasting objects from KONRAD3D system.

Advantage of this approach is the high spatial continuity over vast areas and detection of high-impact weather phenomena, while the disadvantage is that some data have only a qualitative value. But qualitative evaluation could become quantitative by "relaxing" the comparison through neighborhood/thresholding. The link with the nowcasting groups should be strengthened to explore the possible usage of the variables/objects identified through nowcasting algorithms for forecast verification.

1. (a) **Non-conventional data**

The number of applications of non-conventional data grows rapidly. They include:

1. Data from insurances
2. Data from citizens (private met stations, phones), cars
3. Impact data (emergency calls, fire brigade operations) – high spatial resolution
4. Social media (social networks, etc.)
5. Data from cameras and photos

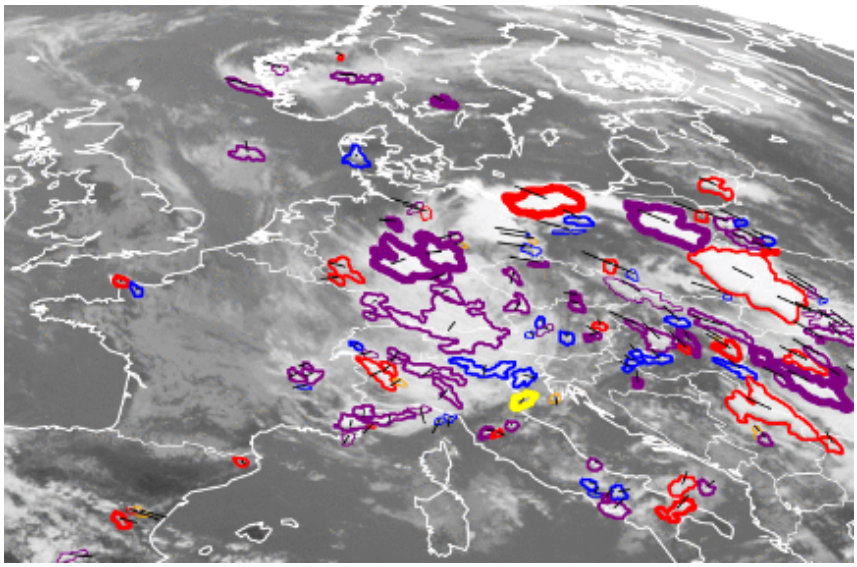


Figure 5: Weather recognition from images [Bin Zhao et al. 2018].

A detailed overview and examples of the studies using new non-traditional sources of data is given in C. Marsigli et al. 2021.

The aim of the Second international verification challenge in 2021 (run by WMO HIWeather Project and Joint Working Group on Forecast Verification Research) was to promote quantitative assessment of high-impact weather, hazards and impacts through the use of non-traditional observations [https://www.emetsoc.org/second-international-verification-challenge/].

Recognition of weather from cameras and photos widely relies on the use of machine learning. For example, in [Bin Zhao et al. 2018], the accuracy of several CNN-RNN Architectures for Multi-Label Weather Recognition from images was studied.

A quantitative estimate of weather variables from images was performed in (Wei-Ta Chu, Xiang-You Zheng, Ding-Shiuan Ding 2017). The average RMSE of temperature estimate was 1.98°C, of humidity, 7.13%, the accuracy of clouds and precipitation estimate was about 76%.

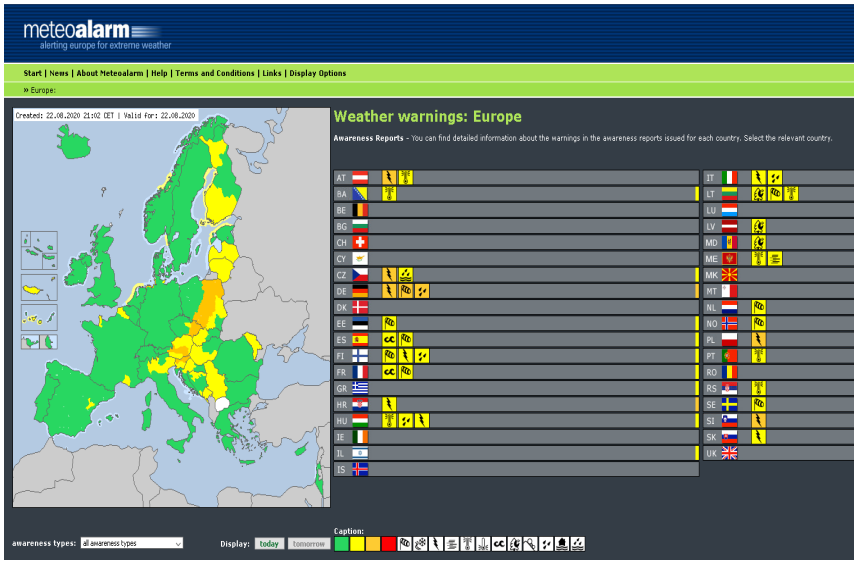


Figure 6: Weather variables determined from photographs [Wei-Ta Chu, Xiang-You Zheng, Ding-Shiuan Ding 2017].

1. (a) **Other data sources**

Other data sources on CW / HIW (mostly storms, but not only) are mostly websites. A universal online resource is the European Severe Weather Database, , operated by European Severe Storms Laboratory.

The information about single event (in general, phenomenon – not only lightning, but generalized HIW event) is presented in a table similar to the one below:

Event	Time and location	Other info/Quality Control
Heavy rain	Inwałd, Małopolskie, Poland (49.87N, 19.39E)<1 km 22-08-2020 (Saturday) 18:30 UTC(+/-15 min.)	based on information from: a report by a weather service, a report on a website, government-based sources/administrative organizations precipitation: 31.2 mm, duration: 0.5 hours Automatic IMWM-NRI weather station measured a rain amount of 31.2 mm in 30 minutes, 26.9 mm in 20 minutes and 20.2 mm in 10 minutes during passage of a thunderstorm. http://monitor.pogodynka.pl/#station/meteo/249190090 Reference: Monitor IMGW, 22 AUG 2020. report status: plausibility check passed (QC0+) contact: ***** ***

Similar information can be obtained from Meteoalarm – Severe Weather Warnings in Europe: https://www.meteoalarm.eu/. By using the dynamic structure of the resource, information about HIW events can be obtained at the spatial resolution level of a few square km, starting from continental, *via* country, to sub-country (city) scale.

Figure 7: Meteoalarm main page

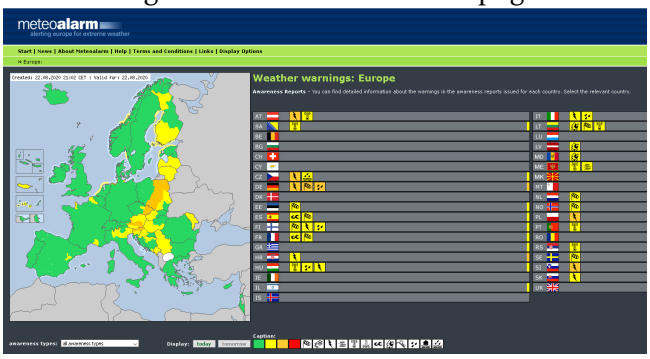


Figure 8: Warnings for selected country

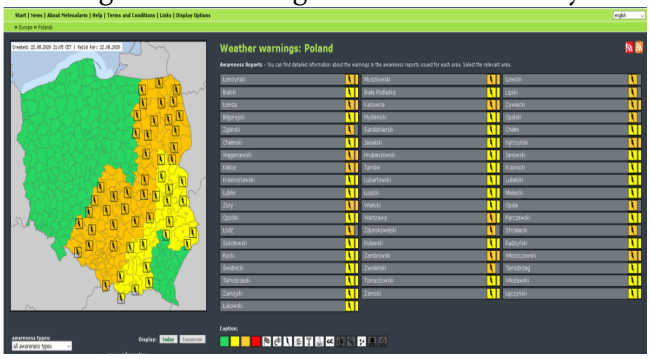


Figure 9: Detailed warning for city/small region

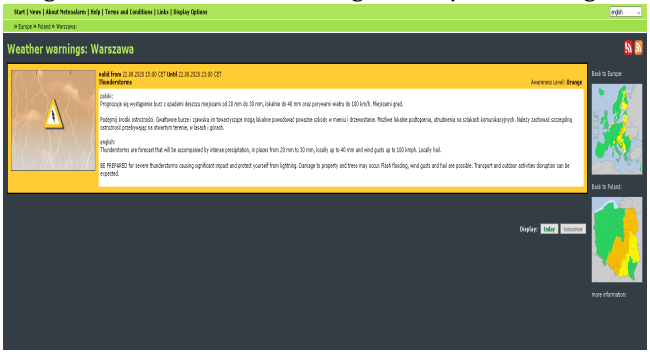
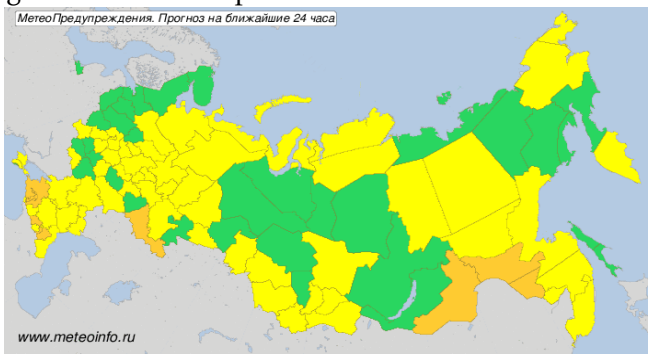


Figure 10: Alert map in Russia similar to Meteoalarm



One important difference is that this portal only allows you to check alerts (forecasts). However, later, for verification, one can compare the data from this webpage with e.g. the data from ESWD/ESSL. For this reason, this webpage should also be considered valuable.

Conclusions

1. Combining all available datasets is usually the best choice
2. The usefulness of data strongly depends on the particular case. For example, during the stormy season, all methods can be equally useful, as well as their combination.
3. For individual cases of thunderstorms, LDN seems to be the best to determine their intensity and location. Supplementing LDN results with radar data would give a full picture of the situation.
4. Data quality and data uncertainty assessment: usage of multiple data sources
5. Introducing uncertainty information in applications – one of the implicit ways: spatial verification methods
6. Closer cooperation with nowcasting, where products for high-impact weather detection are developed

And one final note, definitely written in time and under the influence of the state of the outbreak. In the CoVid-19 era, strangely enough, the number of available data may significantly decrease – as a result of data limitation, e.g. from cancelled flights or sea cruises. That is therefore so important to make the best use of the available data.

References

1. Bin Zhao, Xuelong Li , Xiaoqiang Lu, Zhigang Wang, A CNN-RNN Architecture for Multi-Label Weather Recognition, 2018 Computer Science Neurocomputing
2. Ehrler A., 2018. A methodology for evaluating fog and low stratus with satellite data. Master Thesis. Available at ETH Zürich, Department of Earth Sciences.
3. Gubenko I., PhD thesis, 2016, A study of the physical processes in convective clouds during thunderstorms based on numerical simulation, In Russian.
4. Marsigli, C., Ebert, E., Ashrit, R., Casati, B., Chen, J., Coelho, C. A. S., Dorninger, M., Gilleland, E., Haiden, T., Landman, S., and Mittermaier, M.: Review article: Observations for high-impact weather and their use in verification, Nat. Hazards Earth Syst. Sci., 21, 1297–1312, https://doi.org/10.5194/nhess-21-1297-2021, 2021.
5. Morales G., J. Calvo, C. Román-Cascón and C. Yagüe, 2013. Verification of fog and low cloud simulations using an object oriented method. Poster at the Assembly of the European Geophysical Union (EGU), Vienna (A), 7-12 April.
6. Pardowitz T., 2018. A statistical model to estimate the local vulnerability to severe weather. Nat. Hazards Earth Syst. Sci., 18, 1617–1631.
7. Shishov A.E. A comparison of deep convection detection algorithms based on thresholding techniques applied to Meteosat-11 satellite data for European Russia, Research Activities in Atmospheric and Oceanic Modelling 2021 (Blue Book)
8. Shishov A.E., I.A. Gorlach, An algorithm for the detection and tracking of deep convection using satellite data and integer programming, Hydrometeorological Research and Forecasting, 2020. No. 2. P. 39-59, DOI: https://doi.org/10.37162/2618-9631-2020-2-39-59 (In Russian, abstract in English)

9. Snegurov A.V., Snegurov V.S, Experimental lightning detection network, Russian Main Geophysical Observatory Proceedings, issue 567, pp. 188-200. 2012, In Russian.

10. Tsonevsky, I., C. A. Doswell and H. E. Brooks, 2018. Early warnings of severe convection using the ECMWF extreme forecast index. Wea. Forecasting, 33, 857-871.

11. Wei-Ta Chu, Xiang-You Zheng, Ding-Shiuan Ding Camera as weather sensor: Estimating weather information from single images, Journal of Visual Communication and Image Representation, Volume 46, July 2017, Pages 233-249

12. Westerhuis S., W. Eugster, O. Fuhrer, A. Bott, 2018. Towards an improved representation of radiation fog in the Swiss numerical weather prediction models. Poster presentation at ICCARUS 2018, 26-28 February, DWD, Offenbach (D)

13. Wilson, C. and Mittermaier, M.: Observation needed for verification of additional forecast products, Twelfth Workshop on Meteorological Operational Systems, 2–6 November 2009, Conference Paper, ECMWF, Reading, UK, 2009.

4.2 Approaches to introduce observation uncertainty

Anastasia Bundel, RHM

Quantification of observation uncertainty is important for forecasting all the hydrometeorological variables. For HIW events, which are often the rare ones, it is of extreme importance. Accounting for observation uncertainty can change verification results and it becomes even more important at present when the forecast quality tends to approach the level of observational errors.

Observation uncertainty comes from:

1. Instrumental errors
2. Random errors
3. Reporting errors
4. Representativeness errors
5. Analysis errors
6. Other errors

Observation uncertainty can change verification results. If observation errors are not accounted for during the ensemble verification process, then the investigator may draw inappropriate conclusions about the quality of the prediction system.

Observation errors are the sum of measurement errors and representativeness errors [Janjić et al. 2018]

Methods to account for obs uncertainty [B. Brown, 7th Verification Workshop]

1. Indirect estimation of obs uncertainties through verification approaches (spatial methods, e.g., neighbourhood method with neighbourhood observations). These approaches were explored in AWARE tasks 3.3, 3.4, and 3.6 (see the respective chapters of this report)

2. Incorporation of uncertainty information into verification metrics (e.g., rmse decomposition into components due to “true” frc errors and obs errors)

3. Treat observations as probabilistic / ensembles

4. Assimilation approaches

It was mentioned during the Verification workshop 2020 [!!!] that DA needs background model because cannot deal with missing values, but for verification we can. Verification community might aim for a gridded product without background and accompanied by observation uncertainty mask (reflecting station density, measurement errors, etc.)

The ensemble forecasts run from perturbed initial conditions are perhaps the most traditional way incorporate observation uncertainty. Another method is creating ensemble of observations. One of the most well-known is VERA: Vienna Enhanced Resolution Analysis ensemble [Gorgas, T. and Dorninger, M. 2012]. VERA ensemble has the resolution of 8 km, hourly, 50 members. The generation of perturbations is the core task in the VERA observation ensemble procedure. The observation error information is primarily derived from residuals (i.e. correction increments for individual observations) provided by a data QC scheme for surface station data. The steps towards ensemble analyses are as follows [Dorninger, MesoVICT kick-off meeting 2014]

-Correct station observation values by removing biases derived from deviations proposed by quality control

-Analyse bias-corrected observations = reference analysis

-Generate normal distribution fitted to distribution of quality control outputs

-Create a number of sets of (Gaussian) randomized observation values

-Use perturbed data to create ensemble analyses (Figure 1)

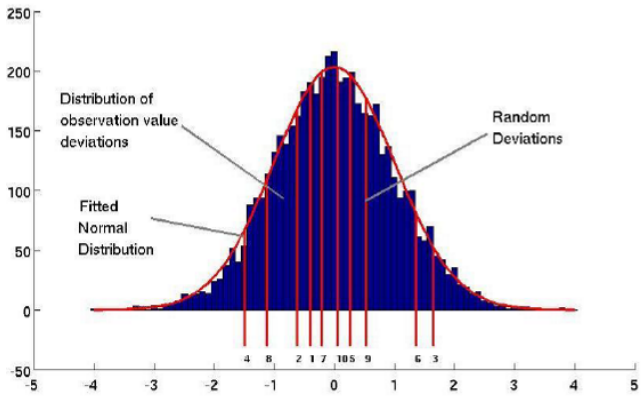


Figure 11: Schematic randomization procedure performed for each station and parameter in VERA analysis.

Such an analysis can be used in the calculation of special verification scores. For example, [Simon Kloiber and Manfred Dorninger 2018] compare the distributions of VERA and model ensembles for the MesoVICT core cases and calculate different scores adapted for observation ensemble, in particular, the CRPS, for different combinations model-observation Ensemble-Deterministic runs.

Radanovics with colleagues studied SAL approach using the observation ensemble [S. Radanovics et al. 2018]. The SAL approach using VERA ensemble was also investigated in the COSMO PP INSPECT by D. Boucouvala [INSPECT Tech. rep. 2019]. In [Ben-Bouallegue et al. 2020], a method is proposed for incorporating representativeness error into the ensemble forecast (The scale mismatch between in situ observations and gridded numerical weather prediction (NWP) forecasts is called representativeness error) [Göber, M., E. Zsoter, and D. Richardson, 2008] Let us note that in [Ben-Bouallegue et al. 2020] the instrumental errors are out of the scope, the representativeness errors being dominant source of error. [Ben-Bouallegue et al. 2020] assess the impact of accounting for observation representativeness on ensemble precipitation verification results. They developed the parametric model of variability on unrepresented scales by fitting a censored, shifted gamma distribution (CSGD). The CSGD is fitted in the form of a conditional distribution for observed precipitation at smaller spatial scale, say B, given the observed precipitation at a larger scale, say A (e.g., the grid scale of an NWP model). They estimated the parameters of CSGD for 24h precipitation accumulations as a function of the averaging spatial scale. They used then the perturbed-ensemble approach that consists of convolving the forecast and observation error distributions [Saetra et al. 2004; Candille and Talagrand 2008]. Each ensemble member gets assigned a random value drawn from the fitted parametric distribution whose scale and shape parameters are a function of the original forecast value: the distribution is centered over the forecast value and its spread accounts for representativeness uncertainty. This approach can also be seen as a forecast downscaling that provides a description of the subgrid-scale uncertainty that is not captured by the NWP model. The additional uncertainty from the perturbed-ensemble approach is merged with the original forecast uncertainty generated by the ensemble system, and together they represent the forecast uncertainty at the observation scale. The verification results showed a large impact of incorporating the representation uncertainty on ensemble scores, in particular at the smaller lead times. As the lead time increases, the forecast error becomes larger predominant. The authors note that more complex approaches cab ne applied, for example, to describe precipitation subgrid variability as a function of the weather situation.

Experiments with incorporating observation uncertainty using MET/METplus verification package

In MET [Newman et al. 2022], a random perturbation approach based on Candille and Talagrand (2008) has been implemented in an attempt to ameliorate the effect of observation errors on the verification of forecasts. The user selects a distribution for the observation error, along with parameters for that distribution. Rescaling and bias correction can also be specified prior to the perturbation. Random draws from the distribution can then be added to either, or both, of the forecast and observed fields, including ensemble members. Details about the effects of the choices on verification statistics should be considered, with many details provided in the literature (e.g. Candille and Talagrand, 2008; Saetra et al., 2004; Santos and Ghelli, 2012). Generally, perturbation makes verification statistics better when applied to ensemble members, and worse when applied to the observations themselves. Normal is the most common choice for observation error. However, the user should realize that with the very large samples typical in NWP, some large outliers will almost certainly be introduced with the perturbation. The lognormal error perturbation prevents measurements of 0 from being perturbed, and applies larger perturbations when measurements are larger. Observation errors differ according to instrument, temporal and spatial representation, and variable type. Unfortunately, many observation errors have not been examined or documented in the literature. Where possible, it is recommended to use the appropriate type and size of perturbation for the observation to prevent spurious results.

In MET Ensemble_Stat tool, (https://met.readthedocs.io/en/latest/Users_Guide/ensemble-

stat.html), the **flag** entry toggles the observation error logic on (**TRUE**) and off (**FALSE**). When the **flag** is **TRUE**, random observation error perturbations are applied to the ensemble member values. No perturbation is applied to the observation values but the bias scale and offset values can be applied to observation values prior to perturbing them. These entries enable bias-correction on the fly. The **dist_type** entry may be set to **NORMAL, LOGNORMAL, EXPONENTIAL,CHISQUARED, GAMMA, UNIFORM**, or **BETA**. The **dist_parm** entry is an array of length 1 or 2 specifying the parameters for the distribution selected in **dist_type**. The **GAMMA, UNIFORM**, and **BETA** distributions are defined by two parameters, specified as a comma-separated list (a,b), whereas all other distributions are defined by a single parameter.

MET can provide a lot of control, enabling the user to define observation error distribution information and bias-correction logic separately for each observation variable name, message type, report type, input report type, instrument type, station ID, range of heights, range of pressure levels, and range of values. This study is our first attempt to incorporate obs uncertainty into verification results. In our setting, the random perturbations for all points in the current verification task are drawn from the same distribution. We have chosen the NORMAL distribution for the air temperature at 2 m, defined by one parameter equal to 1.1 (the option proposed in MET by default for the air temperature at 2 m). Later, we plan to further explore this method, using other distributions. In particular, the gamma or lognormal distribution for precipitation accumulations is of interest.

Below are the first results of our experiments on adding observation perturbations to the ensembles. The model is ICON-EPS-Ru2.2, which is described in more detail in Chapter 3.3 of this report. The observations are SYNOP data on the entire model domain covering the Central Russian federal district. In Figure 2, the scores are shown for one date, 20220223, 00 run. The RMSE and ME are calculated for the ensemble mean. This can explain that there is almost no difference in these scores. The spread is considerably higher for the ensembles with observation error perturbations, and it is closer to the RMSE, which is a desirable property. The difference should be bigger in the probabilistic scores, which are unfortunately calculated only for unperturbed ensembles in MET at present. Figure 3 shows the unperturbed and perturbed ensemble spread aggregated for the test period of 23-28 February 2022. It can be seen from the figure that the difference is large, in particular for the air temperature at 2 m. Our ensembles are found to be under-dispersive overall (Chapter 3.3 of this report). Thus, the products based on the ensemble with observation error perturbations may be useful in better assessing the forecast uncertainty. Further experiments with other distributions to draw the perturbations from are ongoing, in particular, for precipitation, for which the Gamma and Lognormal options exist in MET.

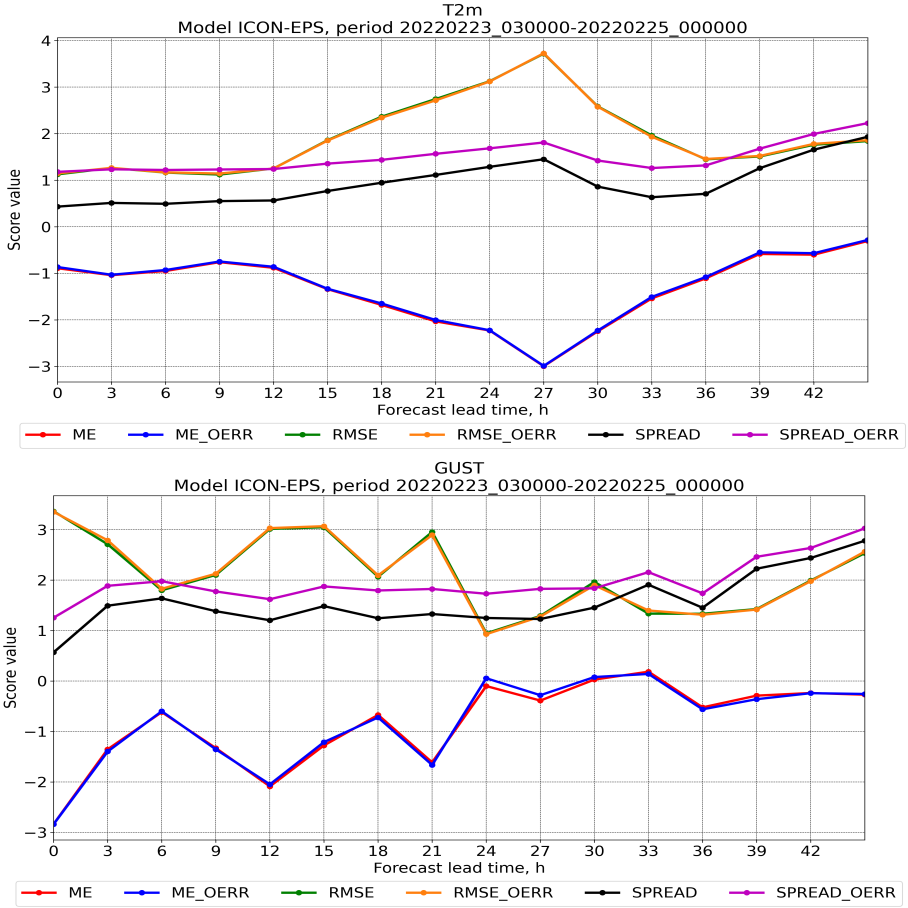


Figure 12: The scores without and with (labelled _OERR) observation error perturbations added to the ICON-EPS-Ru2.2 ensembles, run from 20220223, 00 UTC, (top) air temperature at 2 m, (bottom) wind gusts, Central Russia SYNOP stations. The RMSE and ME are calculated for the ensemble mean.

In [Ben-Bouallegue et al. 2020], it was demonstrated that taking into account observation uncertainty produced larger differences for more high-intensity events in terms of the diagonal elementary skill score (DESS). Thus, accounting for observation uncertainty could be crucial when assessing forecast skill for high-impact events. When the focus is exclusively on extreme events, that is, on the tail rather than on the whole distribution, an accurate estimation of the skill in the presence of observation uncertainty would probably benefit from a more pertinent model definition with the use, for example, of parametric distributions based on extreme value theory (EVT). An example of the application of the EVT to precipitation areas is considered in Chapter 2.3 of this report.

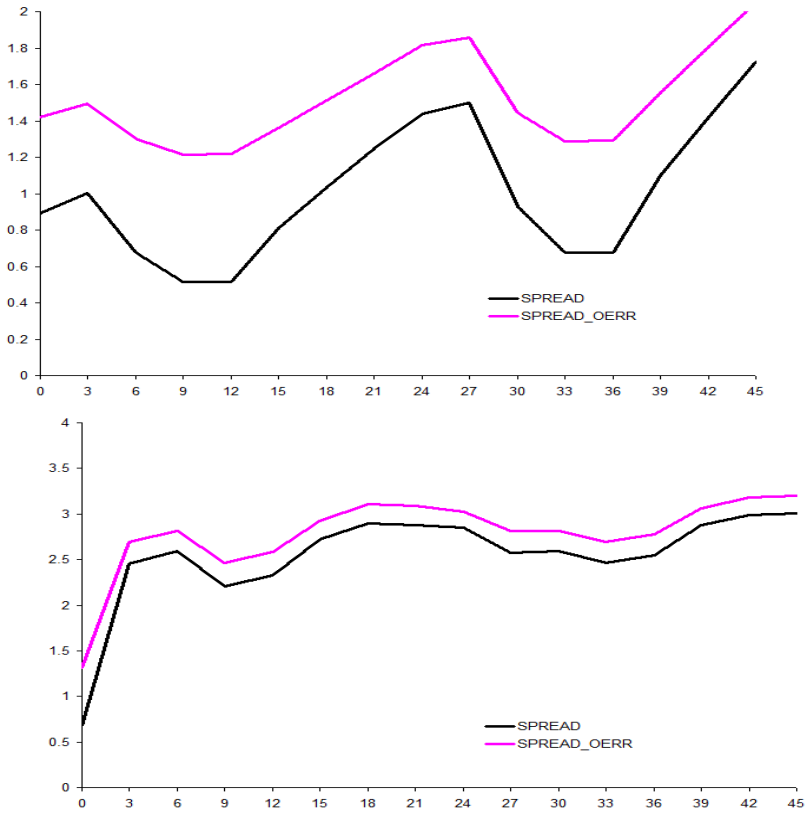


Figure 13: The ensemble spread without (black) and with (purple, labeled _OERR) observation error perturbations added to the ICON-EPS-Ru2.2 ensembles, aggregated for the period of 20220223-20220228, 00 UTC, (top) air temperature at 2 m, K, (bottom) wind gusts, m/s. Central Russia SYNOP stations.

References

1. Bundel A., Gofa F. et al., COSMO Priority Project INSPECT. Final Report. January 2019, 67 . DOI: 10.5676/DWD pub/nwv/cosmo-tr37

2. Candille, G., and O. Talagrand, 2008: Impact of observational error on the validation of ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society* 134: 959-971.

3. Göber, M., E. Zsoter, and D. Richardson, 2008: Could a perfect model ever satisfy a naïve forecaster? Ongrid box mean versus point verification. Meteor. Appl., 15, 359–365, https://doi.org/10.1002/met.78

4. Gorgas, T. and Dorninger, M. 2012. Concepts for a pattern-oriented analysis ensemble based on observational uncertainties. Quarterly Journal of the Royal Meteorological Society, 138:769–784.

5. Janjić, T., and Coauthors, 2018: On the representation error in data assimilation. Quart. J. Roy. Meteor. Soc., 144, 1257–1278, https://doi.org/10.1002/qj.3130.

6. Newman, K., J. Opatz, T. Jensen, J. Prestopnik, H. Soh, L. Goodrich, B. Brown, R. Bullock, J. Halley Gotway, 2022: The MET Version 10.1.2 User's Guide. Developmental Testbed Center. Available at: https://github.com/dtcenter/MET/releases

7. Sabine Radanovics Jean-Philippe Vidal, and Eric Sauquet, 2018, Spatial Verification of Ensemble Precipitation: An Ensemble Version of SAL, Weather and Forecasting, Volume 33: Issue 4, pp 1001–1020, https://doi.org/10.1175/WAF-D-17-0162.1

8. Saetra O., H. Hersbach, J-R Bidlot, D. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Monthly Weather Review* 132: 1487-1501.

9. Santos C. and A. Ghelli, 2012: Observational probability method to assess ensemble precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society* 138: 209-221.

10. Simon Kloiber and Manfred Dorninger, Verification of forecast ensembles by allowing for observation uncertainties in complex terrain, MesoVICT webinar, 11 May 2018.

5 Overview of Appropriate Verification Measures for HIW

Question: How well high-impact weather forecast quality is represented with commonly used verification measures? What is the most appropriate verification approach?

HIW phenomena studied: intense precipitation, thunderstorm (lightning activity)

5.1 Survey for assessment of proper verification of phenomena – continuous vs. discrete verification (occurrence vs. specific values)

Andrzej Mazur, Joanna Linkowska

Institute of Meteorology and Water Management – National Research Institute

Introduction

It can be said that every weather has its impact. Starting with the least inconvenient, like: higher power bills, through moderately troublesome, like: flight delays due to weather conditions, to very dangerous in consequences, like: catastrophes in sea, land and air traffic, destruction caused by a flood or a tornado.

To someone affected, any of these may seem "significant" at that moment. Some impacts are clearly more significant than others. There are four general categories of impacts:

1. Low-impact – minor inconvenience, small and local economic losses, etc.
2. Moderate-impact – minor damage, some social disruption, etc.
3. High-impact – damage, risks to health, broad economic impact, etc.
4. Extreme-impact – dramatic losses, deaths, injuries, major social disruption, etc.

Since every weather has its impact, each weather element can be treated as an impact source. It's just a question of scale and intensity.

1. "regular" elements – temperature, precipitation, wind speed...
2. "specific elements" – visibility limitations, thunderstorms, tornadoes, ...

The verification method may be/could be/should be adapted (and specific) for each element.

Below one can find a list of items done or to be done in this task:

1. Brief research (case studies) to assess applicability of particular method(s);
2. Comparison and judgment whether continuous or discrete methods may/should be applied;
3. Overall final recommendations

Methodology

Survey on (basic) methods applicable to the problem (bold marks jobs done/partially done) consists of:

1. **SAL (Structure/Amplitude/Location) Verification**[‖]
2. **FSS (Fraction Skill Score) verification**[**]
3. **Categorical analysis (Contingency tables and predictands)**

where all the above further on called as "discrete" analysis

1. **Standard evaluation at the grid scale**

hereinafter referred to as "continuous" analysis

1. **Cross- (space-lag) correlation approach and verification**

Structure-Amplitude-Location (SAL) analysis

This approach is defined via three basic elements to be analyzed:

1. **S – structure** – compares the volume of the normalized objects.

The structure component S analyses the size and shape of event objects. The values of S are within [-2,2]. The negative values of S correspond to too small and/or too peaked objects, while positive values indicate too large and/or too flat simulated objects. S=0 indicates a perfect structure.

1. **A – amplitude** – corresponds to the normalized difference of the domain-averaged values

The amplitude component A evaluates the total amount of event occurrence in a predefined region. The values of A are within [-2,2]. Negative values of A correspond to too little and positive values to too much predicted event occurrence, respectively. A=0 denotes perfect forecasts in terms of amplitude.

1. **L – location** – Combinations of a difference of mass centers of fields and averaged distance between the total mass center and individual objects

The location component L quantifies the displacement of observed and simulated precipitation objects, relative to their overall centers of mass. The values of L are within [0,2]. L=0 denotes the perfect value.

Overall, the perfect forecast is expected for S = A = L = 0

The examples of input data for SAL analysis, pertaining to verification of flashrate intensity forecasts and results are shown in the chart of following figures.

[‖]Wernli et al., 2008, SAL – a Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, Mon. Wea. Rev. 136(11), 4470–4487, https://doi.org/10.1175/2008MWR2415.1

[**]Blaylock and Horel, 2020. Comparison of Lightning Forecasts from the High-Resolution Rapid Refresh Model to Geostationary Lightning Mapper Observations, Wea. Forecasting 35, 402-416

The most common case is marked with bold. As it can be seen the parametrization of flashrate intensity based on the CAPE generally overestimates FR compared to the observations.

Fraction Skill Scores (FSS) assessment

This method allows for direct comparison of the forecast and of observed fractional coverage of grid-box events in spatial windows of increasing size. It is supposed to be most sensitive to rare events. Assuming probability of the occurrence of the phenomenon (in the sense of observation) as p_o, and the forecast – p_f , can be defined by the FSS according to the formula below.

$$FSS = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\left(p_f - p_o\right)^2}{\frac{1}{N}\sum_{i=1}^{N}p_f^2 + \frac{1}{N}\sum_{i=1}^{N}p_o^2}$$

with N being number of sub-domains (or windows in an overall domain).

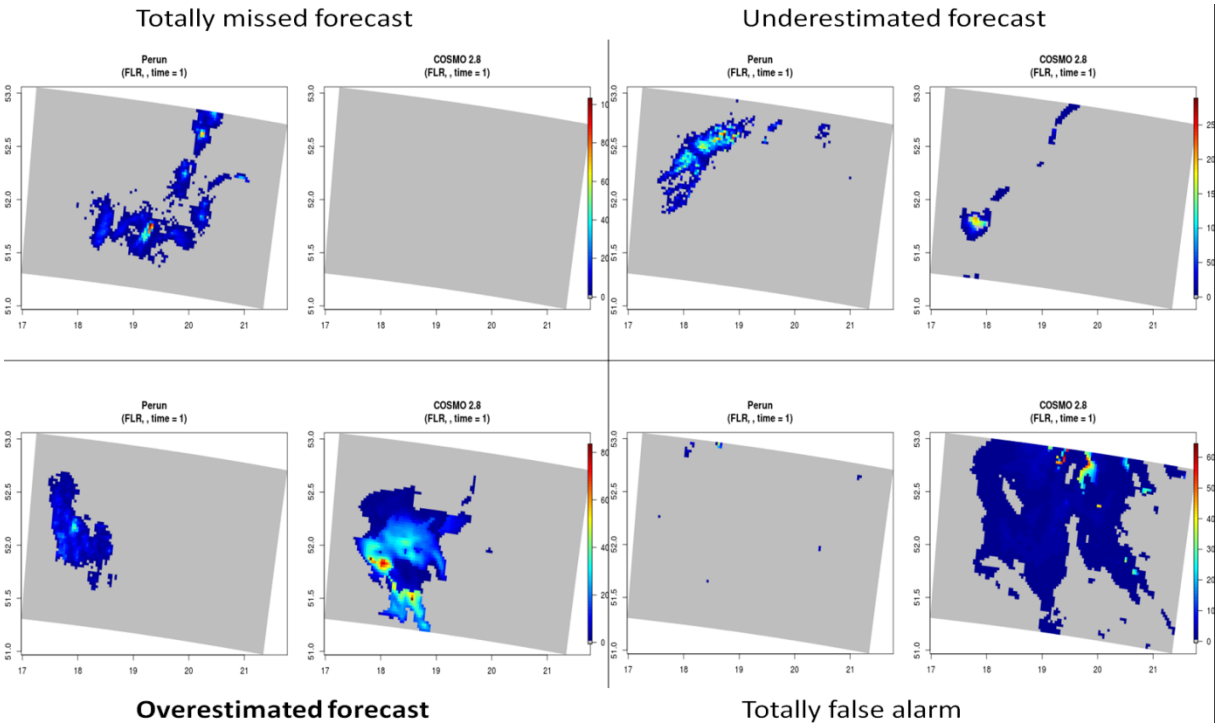


Figure 14: Exemplary verification of flashrate intensity forecasts – Structure-Amplitude-Location approach.

When FSS is equal to 0, there is no correspondence between observations and forecasts. If FSS is equal to 1, it describes a perfect match. Again, exemplary results are shown in the following figures.

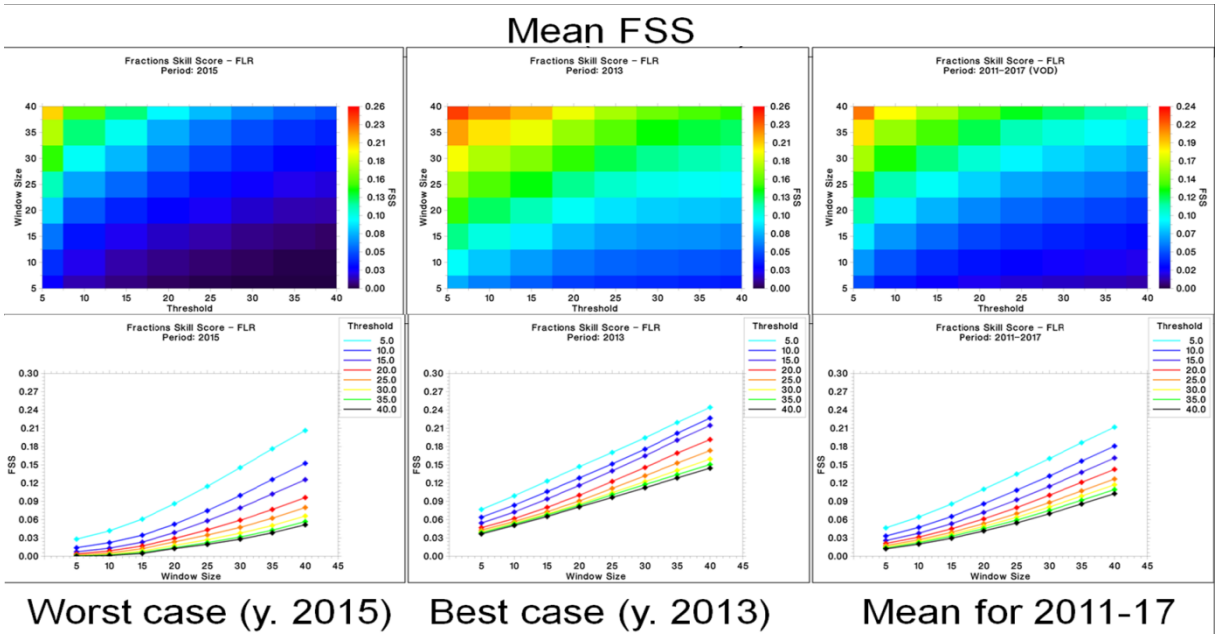


Figure 15: Results of FSS for the worst (2015), the best (2013) year and mean for the entire period of 2011-2017; parametrization of flashrate intensity based on the CAPE.

Categorical analysis based on contingency table:

Forecast given	Event observed	
	Yes	No
Yes	Hit (a)	False alarm (c)
No	Miss (c)	Correct non event (d)

Using values a, b, c and d from the table above, predictands may be constructed as follows:

Predictands used:	**def. n==a+b+c+d**	**range**	**perfect**
Frequency Bias Index	$\frac{a+b}{a+c}$	- to +	1
False Alarm Ratio	$\frac{b}{a+b}$	0 to 1	0
Probability Of Detection	$\frac{a}{a+c}$	0 to 1	1
Probability Of False Detection	$\frac{b}{b+d}$	0 to 1	0
Threat Score	$\frac{a}{a+b+c}$	0 to 1	1
True Skill Statistics	$\frac{a \bullet d - b \bullet c}{(a+c) \bullet (b+d)}$	-1 to 1	1
Equitable Skill Score	$\frac{a-a_r}{(a+b+c-a_r)}$ $a_r = \frac{(a+b) \bullet (a+c)}{n}$	-1/3 to 1	1
Proportion Correct	$\frac{a+d}{(a+b+c+d)}$	0 to 1	1
Success Ratio	$\frac{a}{(a+b)}$	0 to 1	1

Exemplary results are shown in Table 1 and in Fig. 16.

	EQS	FAR	FBI	PFD	POD	SUC	THS	TRS
2012	0.0302	0.8832	2.7196	0.1736	0.2366	0.1169	0.0826	0.0754
2013	0.0773	0.8254	2.4679	0.1483	0.3245	0.1747	0.1249	0.2012
2014	0.0299	0.9060	3.4946	0.1550	0.2193	0.0940	0.0681	0.0935
2015	0.0263	0.8785	2.1706	0.1311	0.1659	0.1215	0.0704	0.0538
2016	0.0555	0.8532	2.7295	0.1592	0.2644	0.1469	0.1030	0.1299
2017	0.0505	0.8296	1.9107	0.1180	0.1981	0.1704	0.0925	0.1002
Mean	0.0420	0.8676	2.3164	0.1499	0.2349	0.1324	0.0898	0.1066

Table 1: Results of contingency tables analysis for the entire period of 2011-2017; parametrization of flashrate intensity based on the CAPE.

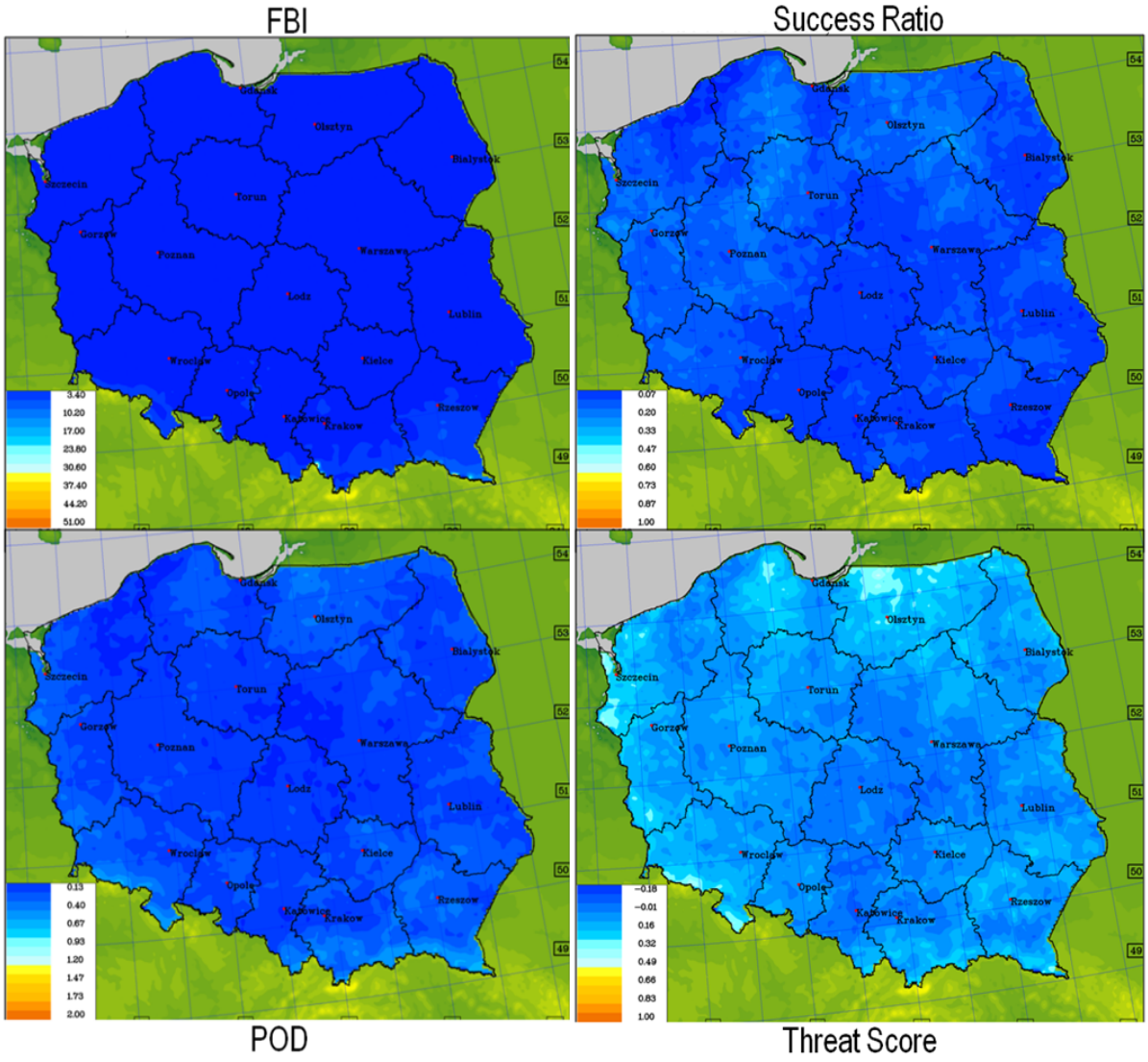


Figure 16: Results of contingency tables analysis for the entire period of 2011-2017 – selected predictands; parametrization of flashrate intensity based on the CAPE.

Standard evaluation at the grid scale ("continuous" analysis)

Year	2011	2012	2013	2014	2015	2016	2017	Mean
ME	2.128	-2.811	-3.674	-3.712	-2.023	-2.291	-1.286	-1.953
MAE	4.712	5.913	2.184	1.516	2.025	3.360	2.817	3.218
RMSE	18.904	18.866	10.556	9.186	11.871	14.695	12.761	13.834

Table 2: ME/MAE/RMSE for consecutive years and mean values for 2011-2017; parametrization of flashrate intensity based on the CAPE.

Continuous analysis requires – in general – the calculation of Mean Error (ME), Mean Absolute Error (MAE) and/or Root Mean Square Error (RMSE). The basic question is – which metric is better? RMSE has the benefit of penalizing large errors more so can be more appropriate in some cases. However, it does not describe average error alone as MAE does. Yet, distinct advantage of RMSE over MAE is that RMSE doesn't use the absolute value – which is good in many mathematical calculations. Results of calculations – both for DMO and for VOD-applied results – are presented in following table and figures. Table 2 contains values of ME/MAE/RMSE for consecutive years and mean values for 2011-2017.

Examples of results for year 2013, 2017 (worse, best) and means for the period are presented in following figures.

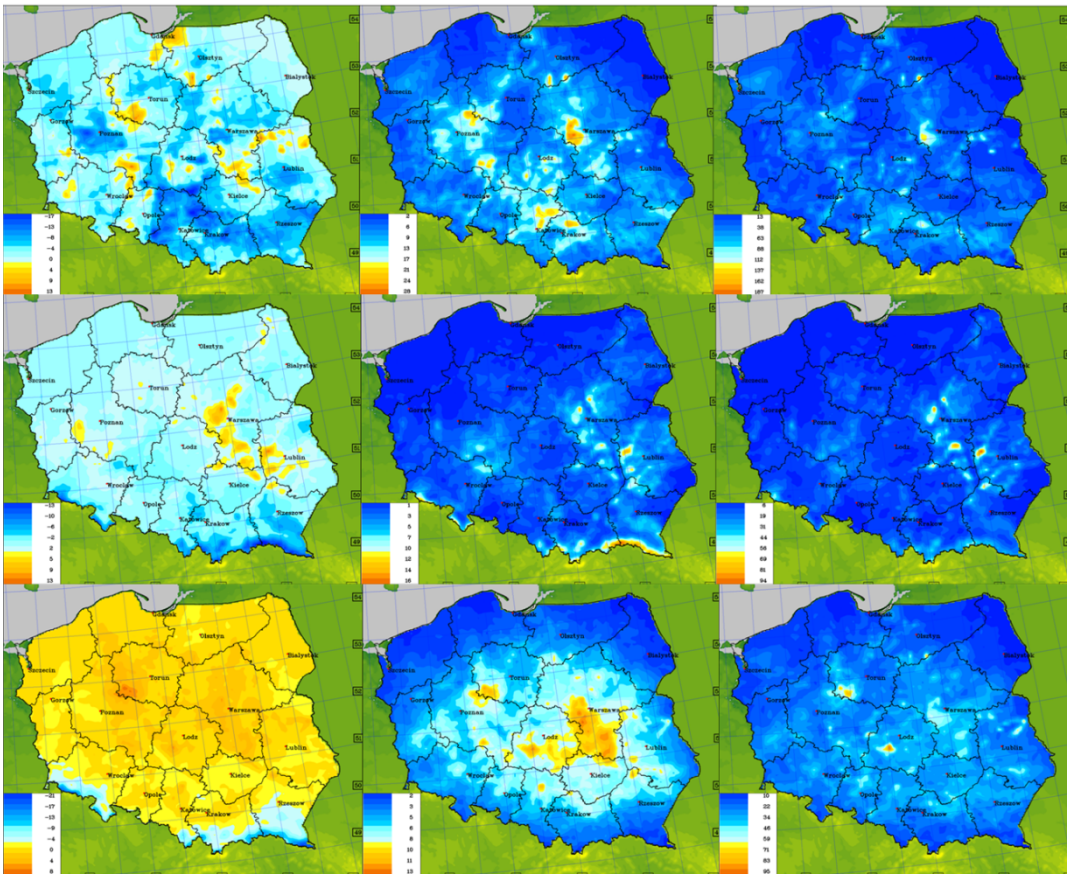


Figure 17: Left to right: ME, MAE and RMSE for 2013, 2017 and mean 2013-2017 as in Table 2.

Space lag (cross-) correlation approach as an addition to basic verification techniques

When overlap the upper left (observations field) and the upper right (forecasts) charts, in most cases they do not match. It is possible to improve the forecast by using the cross-correlation (or space lag correlation) method. To do this (using the example from the figure above) one should:

1. Calculate coordinates of "centres of mass" for both distribution patterns (observations vs. forecasts).
2. Compute vector of displacement (VOD) of forecasts to observations as a difference of the two above.
3. Displace linearly every value of forecasts field by the vector of displacement.

In operational work, VOD is calculated from previous model runs (as compared to observations). It is then assumed to remain constant throughout the next run.

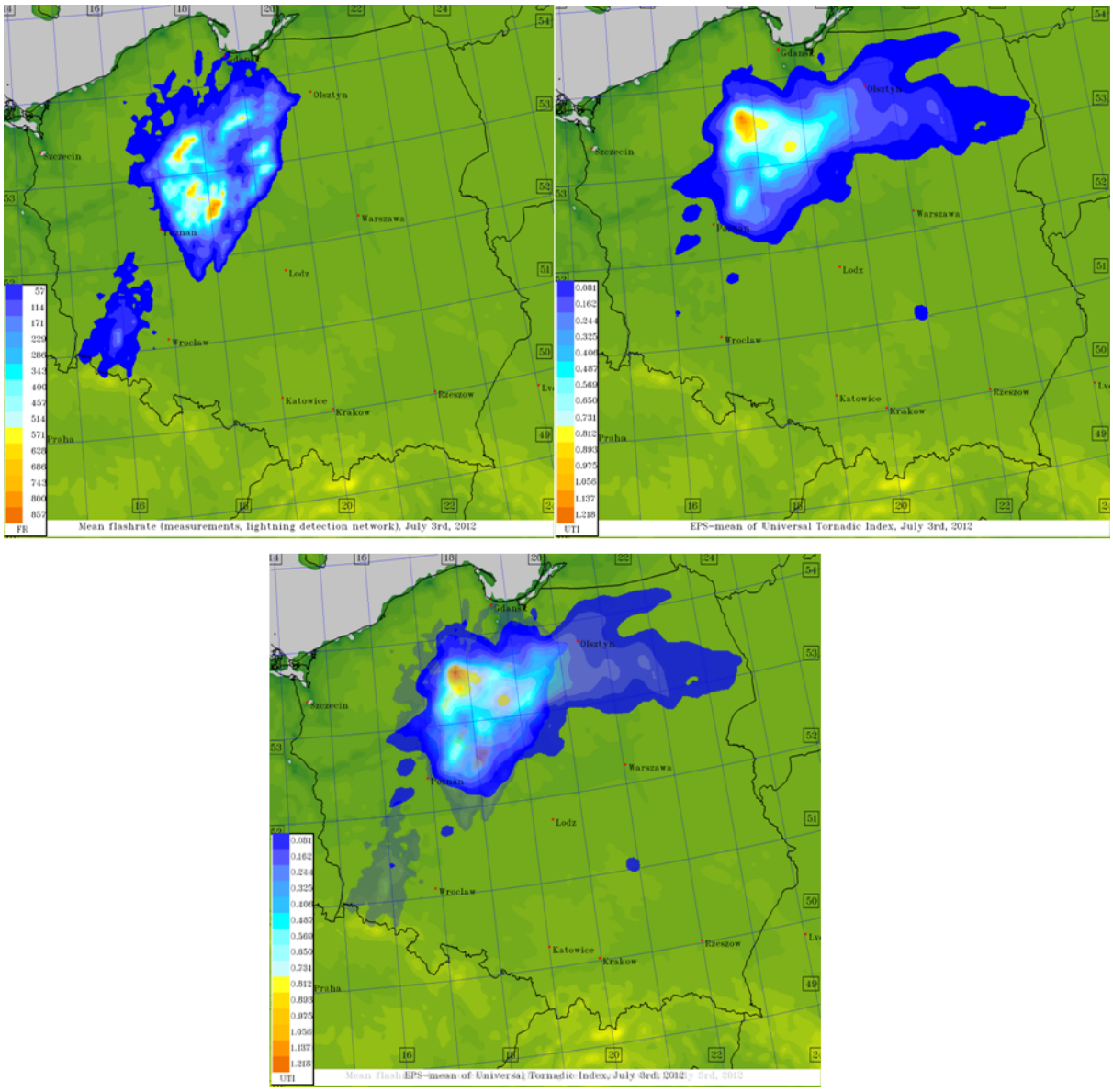


Figure 18: Explanation of VOD procedure – see details in text.

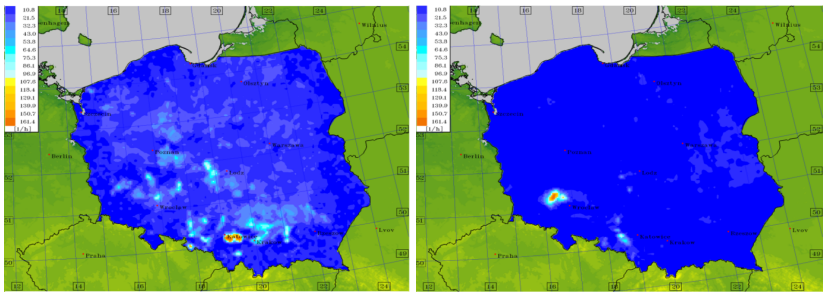


Figure 19: Sample values of (observations – forecasts) for flash rate (lightning frequency). Left - direct model output results, right panel - corrected with VOD procedure.

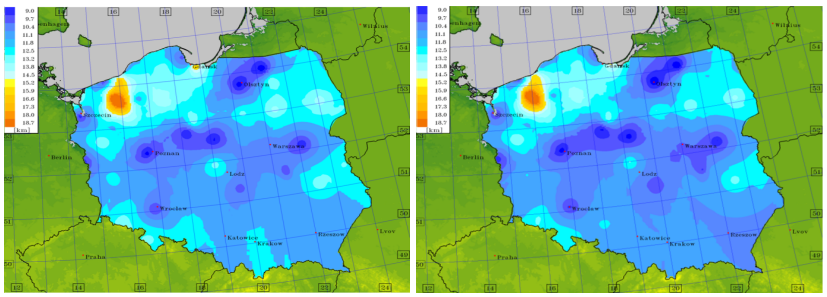


Figure 20: Sample values of (observations – forecasts) for visibility range. Left - direct model output results, right panel - corrected VOD procedure.

All the verification (both "continuous" and "discrete") was done for archive sets of observations (2011-2017). Basic analysis of the results showed that VOD improved virtually all categorical predictands (like FBI, POD, THS...) from 10 up to 45%.

Specific variables

Stability indices

The last part of the report is devoted to specific parameters - stability indicators. These parameters are most often used to summarize the possibility of difficult weather situations. Parameters played an important role in forecasting for more than half a century based on and interpreted upper soundings. The set of these indicators can be considered good prognostic tools as long as the forecasters understand why the values are approaching the critical levels.

Showalter Index (SI)

Historically it was developed for forecasting tornadoes in US, using basic data from radiosondes. It is calculated from the temperature difference of the parcel raised from 850 hPa to 500 hPa.

SI = T500 – $T_{Pcl}500$

Measures the displacement of a parcel raised from the lower to the middle troposphere. It does not take into account the buoyancy (vertical acceleration) above or below 500 hPa, however it takes into account a humidity of 850 hPa when the lifted package reaches saturation, but not above or below 850 hPa, what means that it does not count for an average dryness.

Critical values:

Greater or equal to 0 = stable

-1 to -4 = marginal instability

-5 to -7 = high instability

-8 or less = extreme instability

Total Totals Index (TT)

TT = (T850 - T500) + (T_d850 - T500)

It combines lower tropospheric lapse rate and moisture at low levels; does not account for low level moisture above or below 850 hPa.

Critical values:

Lower than 44 - Convection not likely

44-50 - Likely thunderstorms

51-52 - Isolated severe storms

53-56 - Widely scattered severe storms

Greater than 56 - Scattered severe storms

K Index

This index basically a modification of Total Totals Index for tropical convection; it was intended to forecast convection in US using basic radiosondes data

K Index = (T850 - T500) + (T_d850 – T_{dd}700)

where T_d850 is 850 hPa dewpoint value and T_{dd}700 is 700 hPa dewpoint depression

It combines lower tropospheric lapse rate with amount of moisture in 850-700 hPa layer, but, again, does not account for presence of mid-level dryness. It also does not account for low level moisture others than 850 and 700 hPa. Works best for stations near sea level.

Critical values:

15-25 - small convective potential

26-39 - moderate convective potential

Greater than 40 - High convective potential

SWEAT (Severe Weather Threat) Index

It is in general an evolvement of Total Totals Index, developed to forecast tornadoes and thunderstorms using basic radiosonde data

SWEAT = 12*T_d850 + 20*(TT - 49) + 2*V850 + V500 + 125*{sin[(dd500 - dd850)] + 0.2}

With

T_d850 = 850 hPa dewpoint

TT = Total Totals Index

V850 = 850 hPa wind speed

V500 = 500 hPa wind speed ,

dd500 - dd850 = Directional backing of wind with height (warm advection)

Apart from thermodynamics, it takes account of importance of wind structure and warm advection; does not account for low level moisture above or below 850 hPa, parcel buoyancy or mid-level dryness

Intended for stations near sea level

- If TT less than 49, then that term of the equation is set to zero

- If any term is negative then that term is set to zero

- Winds must be veering with height or that term is set to zero

Does not account for low level moisture above or below 850 hPa, parcel buoyancy or mid-level dryness. Works best for stations near sea level.

Critical values:

150-300 - few severe storms possible

300-400 - severe storms possible

Greater than 400 - tornado possible

Lifted Index (LI)

Mixed Layer (ML) LI describes the difference of temperature of parcel lifted from a layers representing the lowest portion of the atmosphere and the 500 hPa temperature.

LI = T500 – T_{Pcl500}

Measures the buoyancy of a parcel lifted from the lower to the mid-troposphere. Does not account for buoyancy (vertical accelerations) above or below 500 hPa, but accounts for low level moisture implicitly when lifted parcel reaches saturation. It works for stations at most elevations.

Critical values:

0 or greater = stable

-1 to -4 = marginal instability

-5 to -7 = large instability

-8 or less = extreme instability

Convective Available Potential Energy (CAPE)

In general it is an expansion of the Lifted Index, developed to forecast tornadoes and severe thunderstorms.

CAPE = the positive area on a sounding (the area between the parcel and environmental temperature throughout the entire sounding)

It includes no wind information nor information about the strength of the inhibiting convection; can be used to forecast storm intensity, including heavy precipitation, hail, and/or wind gusts, in conjunction with Convective Inhibition (CIN) and Precipitable Water (PW).

Example: maximum vertical motion (without including water loading nor entrainment) can be expressed as $(2*CAPE)^{1/2}$

Critical values:

1 to 1,500 - positive CAPE

1,500 to 2,500 - large CAPE

Greater than 2,500 - CAPE

Convective Inhibition (CIN)

Again, an expansion of the variations of the Lifted Index. Contrary to CAPE, it was developed to forecast non-occurrence of tornadoes and severe thunderstorms.

CIN is the area of the sounding between parcel's starting level and to the level at which CAPE begins to be positive. In this region, the parcel will be cooler than the surrounding environment – thus defining a stable layer.

CIN will be reduced by:

1. daytime heating,
2. synoptic upward forcing,
3. low level convergence,
4. low level warm air advection (especially if accompanied by higher dewpoints).

CIN is most likely to be small in the late afternoon since daytime heating plays a crucial role in reducing it.

Critical values:

0 – 50 - weak Cap

51 – 199 - moderate Cap

Greater than 200 - strong Cap

To sum up – the number of convection indicators is quite large. On the one hand, this is a positive factor, as they collect (and making easier to interpret and understand) the available information about the state of the atmosphere. On the other hand, their results do not always clearly indicate the possibility (or lack of possibility) for the occurrence of the severe weather phenomenon. Moreover, compared to the standard predicted values in the models (temperature, wind, precipitation ...), the possibilities of verification are significantly limited to data from atmospheric surveys. Therefore, it is difficult to satisfactorily define the quality of the forecast of indicators – and hence the possibility of the severe weather phenomenon occurring – over a large area and / or in high spatial resolution.

Conclusions

In the next part of the report (Chapter 6.1), the results for various lighting frequency parameters will be presented as examples of verification of severe weather phenomena. Details will be shown in this study, but it can be stated indisputably that both for long verification periods and for case studies and short-term incidents – if one has the possibility (for variables for which it is possible, of course), should do both discrete and continuous verification. It is because the procedures and results are – for these variables – complimentary.

Conclusions on convection indices remain valid. They should be used as long as there are enough points (i.e., upper air soundings) to verify them.

5.2 Role of SEEPS and EDI-SEDI for the evaluation of extreme precipitation forecasts

Boucouvala D., Gofa F.

Hellenic National Meteorological Service, Hellinikon GR-16777, Athens, Greece

1 Introduction

Precipitation is a parameter highly variable in space and time and exhibits sharp gradients. These characteristics make the evaluation of precipitation forecasts a challenging task which is linked to the observation plurality and spatial inconsistency. On the other hand, there is a large number of possibilities with respect to the choice of score, verification method, spatio-temporal aggregation, which imply different approaches. Most of the verification scores are categorical and based on contingency tables by specifying appropriate thresholds.

Moreover, combining data from a larger number of stations during the evaluation process of NWP forecasts can produce false skill if climatologically diverse regions are combined. In particular, when interest is driven by the presence and implications of heavy precipitation events, one must aggregate regions of similar climatology that will be reflected in the precipitation thresholds that constitute an 'extreme' event in the specific area. Consequently, it is important for HIW events to analyse the relative strengths and weaknesses of commonly used statistical measures but also to highlight the importance of threshold choice especially during the aggregation of results of stations with different climatological characteristics.

This study is focused on the application of two forecast verification skill scores that are related to the geographical and seasonal variations and are already presented in Boucouvala et al. (2016). Short description of the methodology is also given in this paper. The first score is the Stable Equitable Error in the Probability Space (SEEPS) (Rodwell, 2010), which uses the categories "dry", "light precipitation," and "heavy precipitation" based on the climatological cumulative distribution. The second one is the Symmetric External Dependency Index (SEDI) categorical score which is suitable for extreme events as it is equitable, symmetric and does not degenerate for rare events (unlike most categorical scores). It needs however to be adjusted on the climatological characteristics of a specific region by using appropriate thresholds. The combination of these two scores can contribute to the monitoring of model performance and the assistance in the decision making for rare events forecast.

In this study, SEEPS and SEDI scores already applied in the past to assess the predictability of coarser resolution models for the 24 hourly precipitation, are now adjusted (climatologically) and applied for 6h precipitation that is more related to high impact events and are used to evaluate the performance of higher resolution model (COSMO-GR4) and its finer (COSMO-GR1) for all seasons on an annual basis. The objective of this paper through these two metrics is to determine what perspectives these scores provide when climatology is taken into consideration, and focus on forecast assessment of heavy precipitation in order to underline model's ability to reliably capture challenging weather events.

2 Data and Methodology

Statistical Indices: SEEPS is designed to be as insensitive as possible to sampling uncertainty and equitability and adapts to the climate of the region in question. It is based on climatological probabilities of "light" and "heavy" precipitation calculated over a 30-year observations database (1980- 2009) for each station. The station climatology database was

provided by ECMWF while the code calculating SEEPS was developed at Hellenic National Meteorological Service (HNMS) and adapted for this study for 6h accumulated precipitation.

The score involves three categories: 'dry', 'light precipitation' and 'heavy precipitation'. The boundary between the light and heavy categories depends on the relevant climatology for the station at which the score is being calculated. The overall scoring matrix for SEEPS is a function of p_1 (the observed climatological probability of dry weather) and p_2 and p_3 (the observed climatological probabilities of 'light' and 'heavy' precipitation, respectively) at the given observation station (with $p1+p2+p3=1$). Rodwell et al. (2010) assumed p3=p2/2, so the final scoring matrix is the following:

$$\mathbf{S} = \frac{1}{2}\begin{Bmatrix} 0 & \frac{1}{1-p_1} & \frac{4}{1-p_1} \\ \frac{1}{p_1} & 0 & \frac{3}{1-p_1} \\ \frac{1}{p_1}+\frac{3}{2+p_1} & \frac{3}{2+p_1} & 0 \end{Bmatrix}.$$

Threshold between 'dry' and 'light' category is assumed constant at 0.2mm/6h for all time periods and all stations taking into account World Meteorological Organization (WMO) guidelines (Rodwell et al., 2010). Thresholds between 'light' and 'heavy' category are extracted from the database for every station and every month. Therefore, for every month of our dataset, a 3x3 contingency table with the sum of the daily combination of modelled/observed occurrences of each of the 3 categories ('dry', 'light', 'heavy') was computed for each station. The resulting SEEPS index matrix was calculated as the scalar product of the SEEPS weights matrix and the contingency table of total available model/observation pairs for each station averaged over the number of the days of the month. The SEEPS index matrix elements represent the HD (modelled Heavy-observed Dry), LD (modelled Light, observed Dry), LH (modeled Light, observed Heavy), DH (modelled Dry, observed Heavy).

In this study, a weighting distance factor (Rodwell et al., 2010) was also applied in order to avoid over-emphasis of regions with high density. The sum of these components is the total SEEPS value for each month. For our study, the monthly values were also averaged for each season of the whole analyzed period. A perfect forecast has a SEEPS score of 0.

SEDI Symmetric Extremal Dependence Index (Ferro, 2011) is a verification index suitable for low-base (rare) events. It is a function of hit rate (H), and false alarm (F), is complement symmetric, and has a fixed range [-1,1]. It is maximized when H ->1 and F ->0 and minimized when H=0 and F=1. All contingency tables must be non-zero. It is asymptotically equitable, and values >0 imply a forecast that is better than random.

$$SEDI = \frac{\ln F - \ln H + \ln(1-H) - \ln(1-F)}{\ln F + \ln H + \ln(1-H) + \ln(1-F)}$$

Observational and Forecast data: The monthly climatological values of the stations used in this analysis are presented in Fig. 21 (right) and were extracted from the climatological map of Greece (www.climatlas.gr). The complex topography of Greece, which is dominated by both sea and orography, creates variability in both precipitation amounts and frequency, as factors such as elevation, synoptic conditions as well as the region's exposure to wind

lead to small scale climatological patterns (Gofa et al., 2019). A dataset of 6h accumulated precipitation values for 12 months (June 18 to May 2019) were used for 19 stations from various locations (continental, coastal, mountainous) (Fig. 21 left).

With respect to forecast precipitation data, NWP data from operational at HNMS COSMO models were evaluated. Two one-way nested domains were utilized, the coarse domain (4km resolution) covered a wider Mediterranean area, while the inner domain (1km resolution) was set up over the wider geographical domain of Greece. ECMWF operational analysis is used as initial and lateral boundary conditions of the coarse domain.

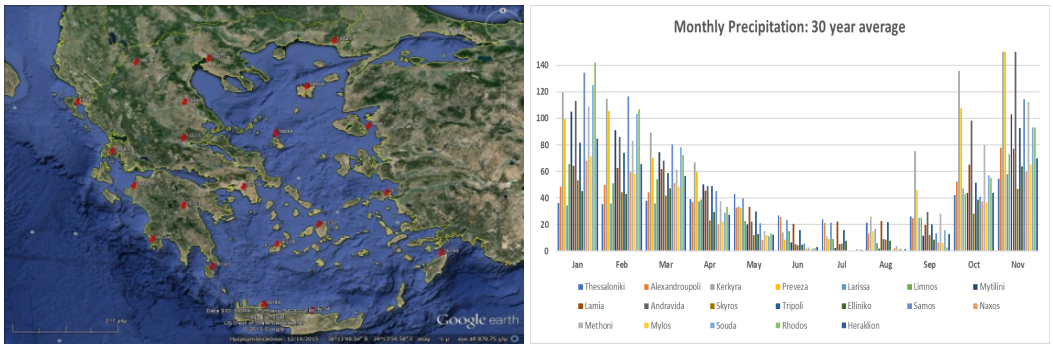


Figure 21: Map of stations that were used for the analysis (left), monthly accumulated precipitation for all used stations (right).

Results

The daily distribution of 6-hourly analysis of precipitation differs for each season as shown for the months of February and June (representative months for DJF and JJA season). In JJA the precipitation in the afternoon is more intense as it is has mainly convective nature, while it is relatively equally distributed in the day in winter period. In addition (not shown), the months with the highest precipitation events were June and January, while the season with no precipitation extremes was MAM for the examined year.

Because of its linearity, the SEEPS score can be broken down into the individual contributions from the six off-diagonal elements of the 3×3 contingency table. This provides some insight into the source of error and also facilitates a comparison of the strengths and weaknesses in model intercomparison. In this study, the emphasis is given on 'Heavy' observed which is related to extreme precipitation events. On a seasonal basis, it is shown that for JJA, the largest SEEPS error contribution comes from predicting the 'dry' category, when 'heavy' was observed. Therefore, summertime heavy precipitation events are significantly underestimated from the model. The study of the 6-hourly precipitation allows us to identify that the maximum error is in the 12-18h interval, when convection mainly occurs in this season. During DJF however, the contribution of HL ('Heavy' observed 'Light' predicted) is the dominant component (purple), so the intense precipitation events are also underestimated but less than in JJA period. In addition, during winter, the daily 6h error distribution exhibits only slightly higher values at night and early morning, a sign of possible underestimation of events at this period of the day. SEEPS values for MAM are the lowest, possibly due to the lack of heavy precipitation events. The differences between COSMO-GR4 and COSMO-GR1 are not so significant on a seasonal basis; therefore COSMO-GR1 results are not shown in this report.

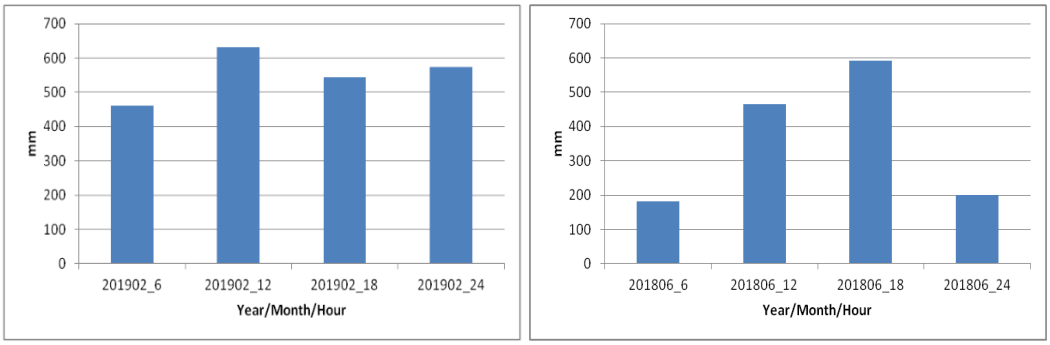


Figure 22: Daily mean 6-h precipitation values for all stations for February (left) and June (right) (hours in UTC).

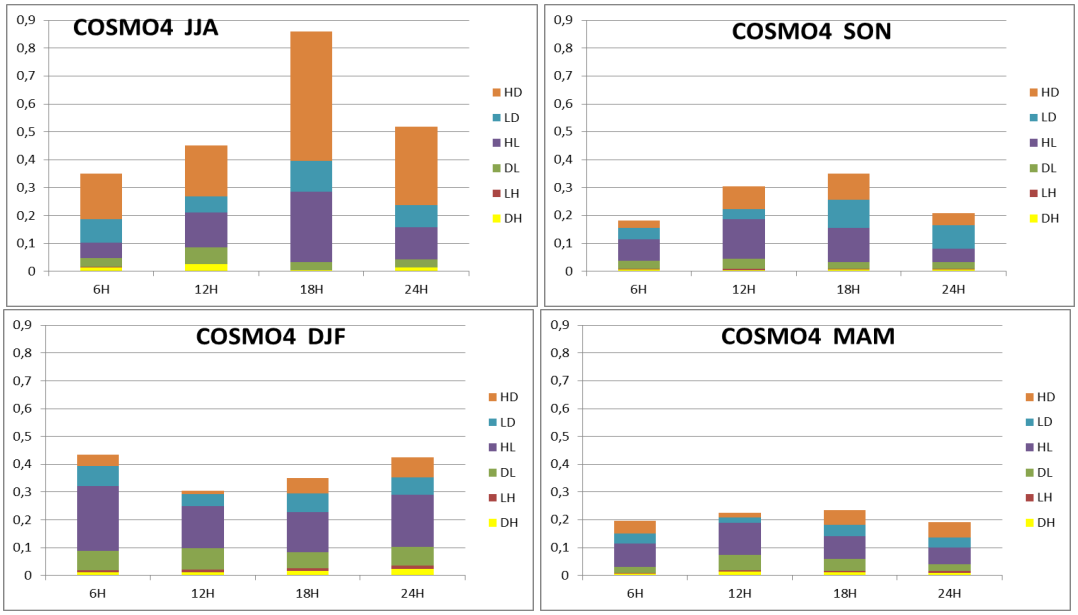


Figure 23: Seasonal SEEPS decomposition on a 6-hourly basis (COSMO-GR4). Colors denote the different components of the index.

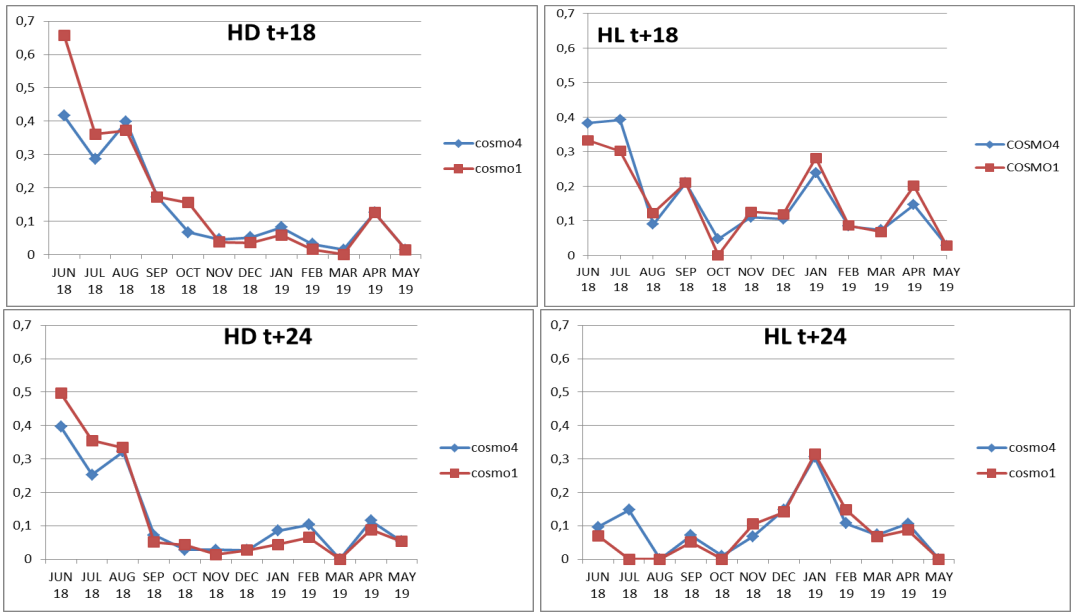


Figure 24: HD and HL components of SEEPS on a monthly basis (COSMO-GR4, COSMO-GR) for 12-18h (upper) and 18-24h (lower).

Monthly graphs for 12-18UTC and 18-24UTC 6h precipitation are also calculated for the SEEPS attributes HD and HL for the whole period. HD (Heavy observed, Dry modelled) is higher in JJA months and drops afterwards. Small secondary maxima are also exhibited in January and April. COSMO-GR1 HD error is slightly higher than that of COSMO-GR4 in JJA. One possible reason is that higher resolution models locate convective precipitation in smaller scale and point verification approach that is used in this methodology, favours the double penalty effect for small spatial misses. The component HL (Heavy observed, Light modelled) is also higher in JJA but only for the 12-18h interval. For this component, COSMO-GR4 values are slightly higher than those of COSMO-GR1, possibly due to the lower predicted values than observed as a result of smoothing related to the lower grid resolution. A secondary significant maximum is shown in January (a month with intense precipitation) implying that in winter, especially during night periods (18-24h), the heavy rain events are underestimated.

SEDI score was also calculated for thresholds based on percentiles-values with low probability to occur (extreme). For example, the 90% percentile value means that according to climatology there is 5% chance that precipitation higher than this occurs. This threshold-based approach is more suitable when stations of different climatology are taken into account for the extraction of average scores. The monthly percentile values for each station were extracted from the 30-year database that was mentioned earlier. SEDI values for 6-12UTC and 18-24UTC intervals are plotted for each season for COSMO-GR4.

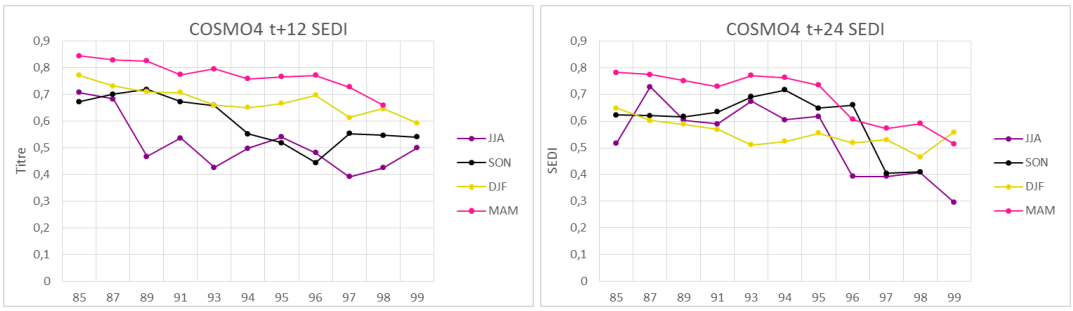


Figure 25: Seasonal SEDI index for each season for COSMO-GR4 for 06-12UTC (left) and 18-24UTC (right) intervals.

SEDI score values (best is 1), generally reduce with increasing percentile values especially for 18-24UTC precipitation. Worse SEDI values during daytime are worse in JJA season, while score is improving in MAM. This result is consistent with the analysis when SEEPS index was considered for the same season. Moreover, SEDI score values for DJF nighttime period, are worse than daytime and this also confirms what was previously found for SEEPS score.

Conclusions

In this study, effort was given to include in the evaluation process of NWP precipitation forecasts, the aspect of climatology by making regions comparable using variable thresholds depending on precipitation climatology. SEEPS and SEDI scores were adjusted and applied on 6h precipitation intervals, as the focus was on the model's ability to capture in a timely manner intense precipitation events. SEEPS is based on a 3×3 contingency table and measures the ability of a forecast to discriminate between 'dry', 'light precipitation', and 'heavy precipitation', while SEDI is a verification index suitable for low-base (rare) events.

The analysis of one-year period allowed to identify the source of forecast errors for two high resolution models (COSMO-GR4 and COSMO-GR1) on a seasonal and monthly basis. The methodology that was developed, reveals the relative contribution and source of error of each model. Furthermore, it permits a more fair evaluation of forecast performance during intense precipitation events, when a model domain of variable climatology is considered. Climatologically-derived and site-specific percentile thresholds, combined with large time-windows, give large enough sample to make SEDI and SEEPS robust and informative, both suggesting that the higher resolution model is more capable (in most cases) to represent high intensity precipitation events.

References

1. Boucouvala D, Gofa F, Fragkouli P (2017) Complimentary Assessment of Forecast
2. Performance with Climatologically Based Approaches. In: Karacostas T., Bais A.,
3. Nastos P. (eds) Perspectives on Atmospheric Sciences. Springer Atmospheric Sciences. Springer, Cham. https://doi.org/10.1007/978-3-319-35095-0_107
4. Ferro CAT, Stephenson DB (2011) Extremal dependence indices improved verification measures for deterministic forecasts of rare events. Weather Forecast 26: 699-713.
5. Gofa F, Mamara A.; Anadranistakis M, Flocas H (2019) Developing Gridded Climate

Data Sets of Precipitation for Greece Based on Homogenized Time Series. *Climate 7*: 68. https://doi.org/10.3390/cli7050068

6. Rodwell MJ, Richardson DS, Hewson TD, Haiden T (2010): A new equitable score suitable for verifying precipitation in NWP. Q. J. R. Meteorol. Soc.: 136, 1344-1363.

5.3 Extreme Value Theory (EVT) approach- Fitting precipitation object characteristics to different distributions

A. Muraviev, A. Bundel, RHM

Papers published within this research:

A.V. Muravev, A.Yu. Bundel, D.B. Kiktev, A.V. Smirnov, Expertise in spatial verification of radar precipitation nowcasting: identification and statistics of objects, situations and conditional samples // Hydrometeorological Research and Forecasting. 2022, 2, 384, pp. 6-52, DOI: https://doi.org/10.37162/2618-9631-2022-2-6-52 [In Russian, abstract and figures in English available]

A.V. Muravev, A.Yu. Bundel, D.B. Kiktev, A.V. Smirnov, Verification of radar precipitation nowcasting of significant areas using the generalized Pareto distribution. Part 1: Elements of theory and methods for estimating parameters // Hydrometeorological Research and Forecasting, 2022, 3, 385, pp. 6-41, DOI: https://doi.org/10.37162/2618-9631-2022-3-6-41 [In Russian, abstract and figures in English available]

A.V. Muravev, A.Yu. Bundel, D.B. Kiktev, A.V. Smirnov, Verification of radar precipitation nowcasting of significant areas using generalized Pareto distribution. Part 2: application to forecasts in warm and cold periods of 2017–2018 // Hydrometeorological Research and Forecasting, 2022, 3, 385, pp. 42-77, DOI: https://doi.org/10.37162/2618-9631-2022-3-42-77 [In Russian, abstract and figures in English available]

The study considers the problems of modeling the extreme values on the example of contiguous precipitation areas observed and predicted by the radar-based precipitation nowcasting system of the Hydrometcentre of Russia. Precipitation fields were converted into objects using spatial averaging (9 nearest points) and the isoline of 1 mm/h. The sets of objects sizes (or areas) exceeding certain area thresholds were formed so that they at least partially satisfy the conditions of physical and statistical independence for applying the extreme value theory (EVT). The model of "peaks above the threshold" described by the generalized Pareto distribution is chosen as the basic model of extreme values.

Analysis setup:

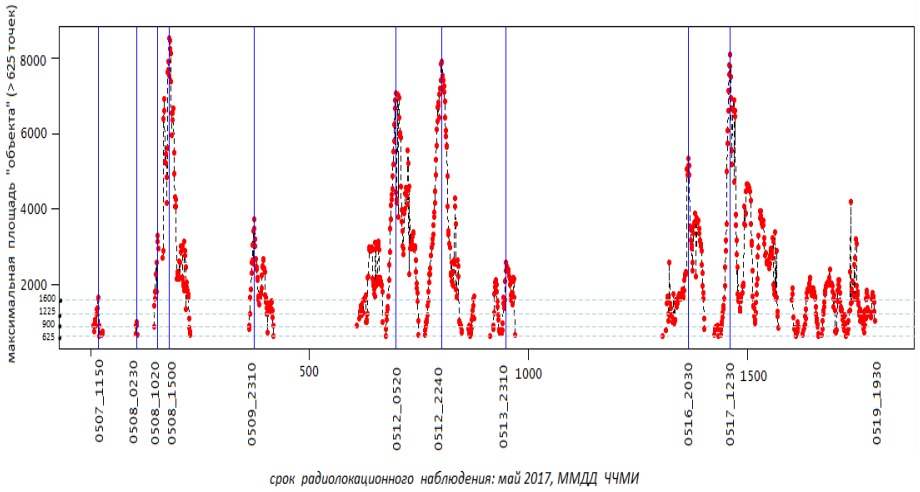
1. The core of the nowcasting system is the statistical STEPS scheme (Short Term Ensemble Prediction System) (*Seed 2003*, *Seed 2004*, *Bowler N. et al., 2006*)
2. Verification period: Warm: May-September 2017

Cold: November 2017-March 2018

1. Nine DMRL-S radars in the Central Federal District of Russia were used as reference data
2. 10 min time step until 3 h
3. Grid size of about 2 km, 256x256 grid points domain

The computational procedures were performed using the tools and graphical representation available in the R language mainly. Objects were selected using the mathematical module FeatureFinder() of the SpatialVx library. To estimate the distribution parameters, we selected objects with sizes of at least 25×25, 30×30, 35×35, and 40×40 points in

a two-kilometer grid. Figure 1 shows the object areas of not less than 625 grid points in subsequent observation fields, the blue lines indicating the times of maximum object area occurrence in each precipitation situation.



The generalized Pareto distribution (GPD) was used with fixed location thresholds (Pareto thresholds) equal to the selected object sizes (625, 900, 1225 and 1600 points). The GPD is an approximation of the peaks-over-threshold distribution, where the peaks are taken from the data of the generalized extreme value distribution. The times of the peaks on the time axis being the Poisson point process is the sufficient condition for satisfying the conditions of the first extreme value theorem (the Fisher–Tippett–Gnedenko theorem). It was checked (using the R Poisson package) that the times of the peaks in our data satisfy the conditions of the Poisson point process. Therefore, it justifies applying the GPD to precipitation object area maxima.

The GPD parameters were estimated using 1) maximum likelihood methods, 2) maximum likelihood, 3) L-moments, and 3) Bayes with stochastic Markov chain modeling. Based on the shape estimates and their confidence intervals, it can be argued that all four methods led to consistent conclusions about the shape parameter for the threshold of 625 grid points. The standard method is undoubtedly the maximum likelihood method (MLE) associated with a modified Chi-square minimum method [*Cramer 1999*], which, as mentioned above, can also replace the Akaike information criterion under some general assumptions. However, on small samples, MLE can lead to unnatural parameter estimates, which has led to the suggestion of a truncated Bayesian correction (GMLE) [*Martins, E. S. and Stedinger, J. R.* 2000, 2001]. The L-moments methods are attractive due to the simplicity of calculations and the statistical robustness of the estimates. However, [in *Martins, E. S. and Stedinger, J. R.* 2001], statistical experiments showed the advantage of the GMLE method over the L-moments for samples of medium size and heavy tails, i.e. when the shape parameter is at least positive. Full confidence in the Bayesian parameter estimation strategy is hindered by a lack of experience in the broad sense, including insufficient mastery of the methodology, and experience in applying this strategy to extreme values in particular. The existence of many methods for estimating parameters confirms, on the one hand, the complexity of the statistical analysis of extrema, and, on the other hand, excludes the existence of one general and universally applicable method. We used the GMLE method to bring the estimates of the nowcasting quality to a small number of observable results.

We studied the modeling quality for all thresholds using histograms and Chi-square crite-

rion of the quality of approximation with Generalized Pareto distribution (GPD). Figures 2 and 3 show the histograms of the size distribution of objects no smaller than 625, 900, 1225, and 1600 points and the GPD density values connected by linear segments, for one radar. Parameter estimation method is GMLE. The titles of the panels indicate the sample sizes, number of gradations automatically calculated, extreme values and median sizes, as well as the parameter estimates. Differences and similarities between the histograms and approximating Pareto distribution density curves are visually visible for objects with sizes above the thresholds of 625, 900, 1225 and 1600 points. Let us note that, on average, the number of objects in the forecast fields is bigger than in the observation fields in our sample. Since a larger Pareto threshold selects a subset of the maxima selected for a smaller threshold, the approximation of the higher-threshold subset by the Pareto distribution must increase the scale (going to the right along the tail) and change the shape.

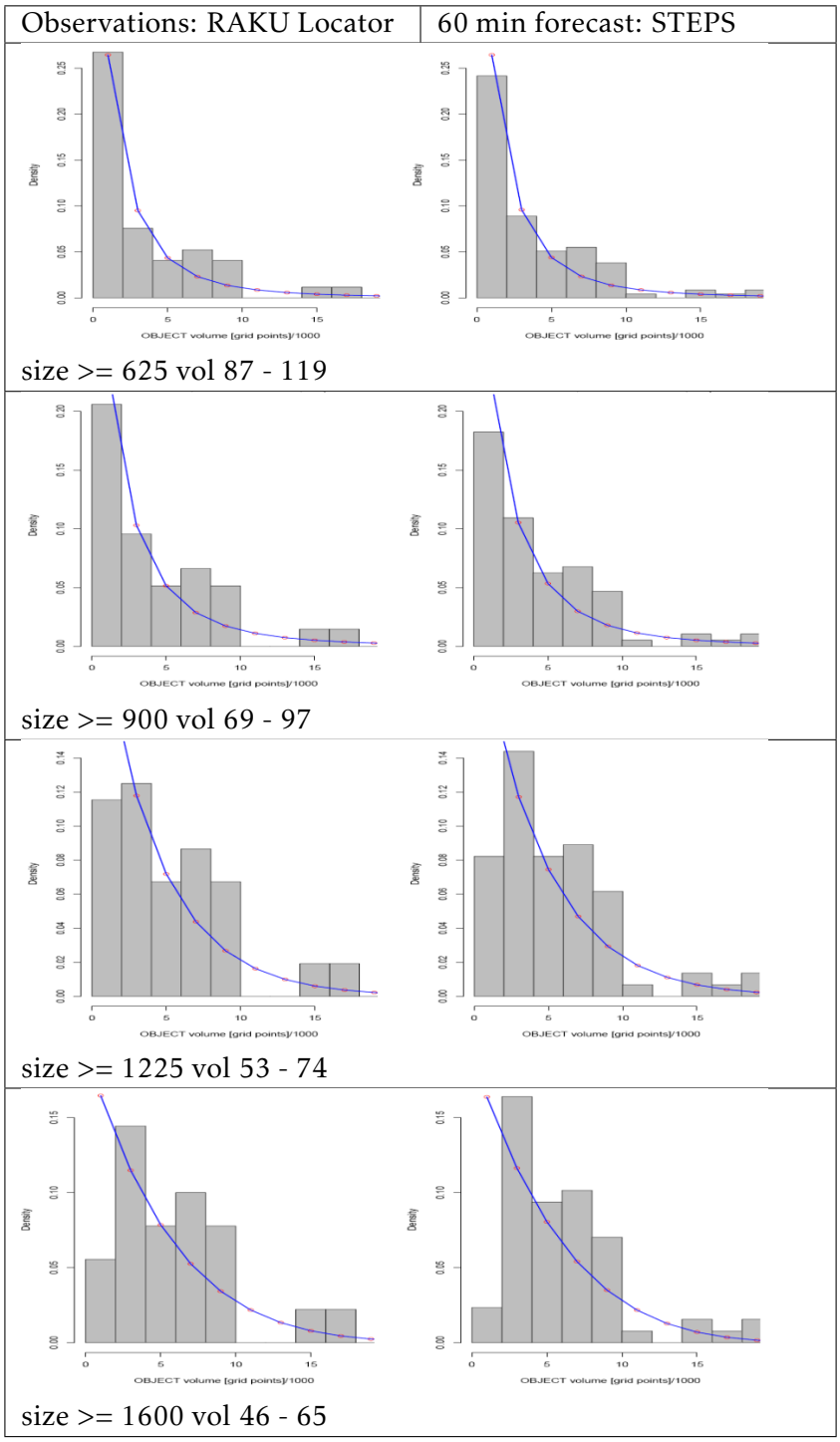


Figure 26: **Warm period**. Histograms and Pareto distribution approximation of object sizes in precipitation fields in Kursk radar observations (RAKU, left column) and forecasts (STEPS-60, right column) for 60 min. The Pareto threshold is (from top to bottom) 625, 900, 1225 and 1600 field points. Dimensions are given in size/1000 scale.

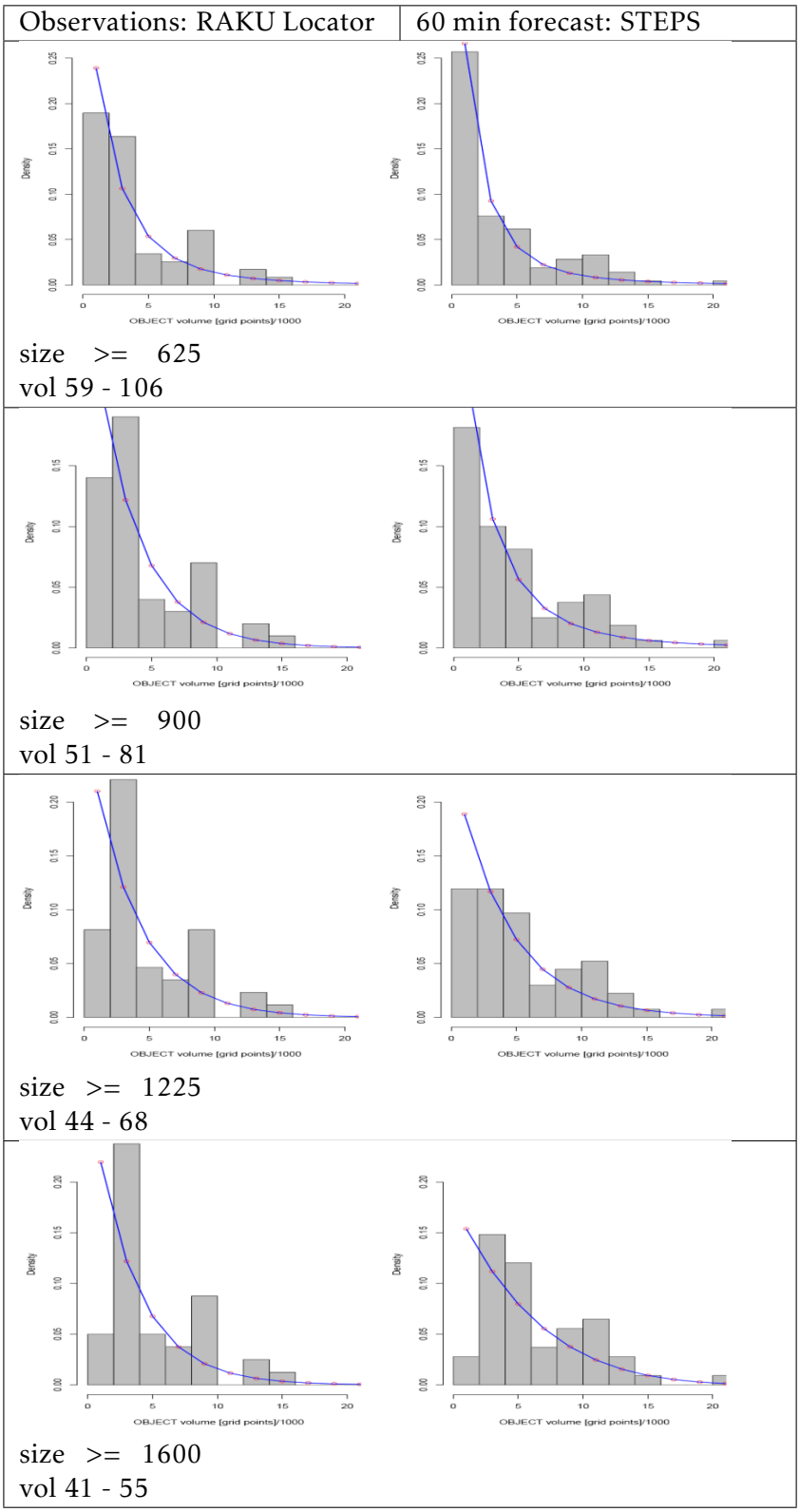


Figure 27: **Cold period**. Histograms and Pareto distribution approximation of object sizes in precipitation fields in Kursk radar observations (RAKU, left column) and forecasts (STEPS-60, right column) for 60 min. The Pareto threshold is (from top to bottom) 625, 900, 1225 and 1600 field points. Dimensions are given in size/1000 scale.

	WARM PERIOD							
	RADAR				STEPS-60 MIN			
RADAR(cases; ndeg)	625	900	1225	1600	625	900	1225	1600
RAKU (87-46;11)	**13.190**	**13.106**	**13.420**	**14.009**	**14.721**	**14.201**	**15.788**	**24.245**
RATL (80-57;11)	**4.097**	**5.674**	**8.708**	**12.476**	**11.395**	**12.140**	**10.565**	**17.086**
RAVO (90-65;11)	**6.921**	**8.417**	**13.148**	**19.733**	**4.637**	**7.423**	**11.288**	**21.318**
RUDB (77-52;11)	**6.690**	**9.289**	**13.859**	**21.479**	**7.695**	**10.024**	**15.256**	**21.446**
RUDK (97-66;10)	**6.862**	**8.432**	**13.818**	**18.803**	**5.947**	**8.044**	**13.332**	**20.265**
RUDL (93-61;12)	**4.711**	**6.310**	**9.959**	**13.958**	**14.335**	**16.637**	**22.859**	**30.741**
RUDN (85-61;13)	**4.301**	**5.301**	**9.213**	**12.080**	**7.474**	**10.088**	**14.666**	**22.267**
RUWJ (86-61;11)	**4.059**	**4.499**	**9.204**	**12.638**	**5.220**	**7.727**	**13.841**	**21.949**
	COLD PERIOD							
	RADAR				STEPS-60 MIN			
RADAR	625	900	1225	1600	625	900	1225	1600
RAKU (59-41;12)	**10.849**	**12.427**	**16.501**	**22.069**	**13.646**	**10.810**	**10.982**	**15.125**
RATL (48-25;13)	**10.775**	**10.747**	**11.624**	**11.623**	**13.599**	**12.759**	**10.450**	**7.272**
RAVO (54-29;10)	**10.296**	**11.282**	**15.997**	**23.012**	**8.947**	**9.096**	**15.831**	**23.653**
RUDB (46-25;10)	**7.617**	**8.678**	**11.632**	**9.365**	**7.689**	**7.192**	**10.107**	**14.690**
RUDK (41-18;10)	**11.502**	**13.442**	**13.641**	**16.047**	**12.697**	**12.139**	**15.109**	**22.009**
RUDL (47-25;7)	**3.604**	**3.020**	**2.056**	**1.888**	**12.928**	**4.424**	**3.651**	**2.909**
RUDN (41-20;11)	**12.630**	**11.117**	**10.904**	**12.481**	**6.423**	**5.255**	**7.355**	**10.534**
RUWJ (27-16;12)	**8.370**	**10.298**	**14.593**	**24.938**	**13.222**	**12.296**	**13.484**	**18.493**

Table 3: Chi-square test values for assessing the quality of histogram approximation by the Generalized Pareto distribution with estimated scale and shape parameters at thresholds of 625, 900, 1225, and 1600 points.

Table 3 summarizes Chi-square estimates for object areas in radar fields (RADAR columns) and in 60 min forecast fields (STEPS-60 MIN columns) for tests in warm and cold periods of the year. In the R hist() function, the number k of bins (necessary to calculate the Chi-square criterion) is determined using *the Sturges rule* : $k = 1 + [\lg_2(n)]$, where n is the sample size. The analysis was carried out for each period and for each observation-forecast pair.

Note: In the RADAR column, next to the identifier, in brackets are indicated ("sample size of observations" for a threshold of 625 points - for a threshold of 1600 points; an estimate of the number of degrees of freedom).

The yellow background highlights the values that exclude the Pareto distribution from the set of suitable approximations; on the histograms (e.g., Figure 26 and 27), this is, as a rule, the second bin being larger than the first one, that is, the violation of the characteristic Pareto distribution density curve. The values highlighted in red reflect one of the most important conditions of the second extreme value theorem, *threshold stability*: the larger the threshold, the more accurately the data is modeled by the Pareto distribution. However, in real samples of a limited size rapidly decreasing with increasing threshold, this phenomenon should be recognized as a rare success. Thus, to draw the conclusions, we chose two basic Pareto thresholds, 625 and 900 points, as these thresholds provide the best quality of approximation of the data with the GPD (Table 3).

We further introduce a special metric for estimating the STEPS quality. The standard errors of estimates of GPD parameters are used to construct 95% confidence intervals (CI) and to subsequently compare estimates of the scale and shape parameters based on the intersec-

tion ratio (IR). The boundaries of the confidence interval are determined in a standard way (estimate ± 1.96 * error). Let's write lower and upper limits of the confidence intervals like (L_1 , U_1) and (L_2 , U_2), respectively. The intersection ratio (IR), visually obvious, is determined as follows:

IR = (min (U_1 , U_2) - max (L_1, L_2))/(max(U_1 , U_2) - min(L_1 , L_2))

The intersection ratio gives a diagnostic estimate of model ability to reproduce vast contiguous precipitation areas (or other extremes). The intersection ratios (IR) of confidence intervals for estimates of the scale and shape parameters are summarized in Table 2. Let us empirically choose a level of "failure", e.g., intersect <50%, and mark it in red.

RADAR		IR (%) SCALE				IR (%) SHAPE			
		warm period		cold period		warm period		cold period	
	threshold\ lead time	625	900	625	900	625	900	625	900
RAKU	30	80	74	75	68	84 (+ +)	74 (+ +)	78 (+ +)	47 (0 +)
	60	83	77	50	63	85 (+ +)	76 (+ +)	68 (+ +)	44 (0 +)
	90	83	73	23	38	83 (+ +)	76 (+ +)	54 (+ +)	33 (0 +)
	120	79	68	21	20	80 (+ +)	72 (+ +)	54 (+ +)	23 (0 +)
RATL	30	39	27	62	78	74 (+ +)	41 (0 +)	78 (+ +)	87 (+ +)
	60	48	23	52	55	72 (+ +)	36 (0 +)	69 (+ +)	71 (+ +)
	90	35	17	47	53	66 (+ +)	32 (0 +)	67 (+ +)	69 (+ +)
	120	34	19	50	62	65 (+ +)	34 (0 +)	68 (+ +)	73 (+ +)
RAVO	30	54	38	70	76	78 (+ +)	40 (0 +)	76 (+ +)	71 (+ +)
	60	58	37	64	74	76 (+ +)	36 (0 +)	72 (+ +)	67 (+ +)
	90	56	46	61	63	77 (+ +)	38 (0 +)	66 (+ +)	63 (+ +)
	120	52	35	50	64	69 (+ +)	34 (0 +)	64 (+ +)	60 (+ +)
RUDB	30	92	88	75	78	93 (+ +)	91 (+ +)	72 (+ +)	72 (+ +)
	60	87	90	48	67	91 (+ +)	92 (+ +)	60 (+ +)	64 (+ +)
	90	81	94	46	56	92 (+ +)	94 (+ +)	56 (+ +)	57 (+ +)
	120	88	92	44	56	89 (+ +)	92 (+ +)	55 (+ +)	56 (+ +)
RUDK	30	78	85	69	73	80 (+ +)	79 (+ +)	75 (+ +)	70 (+ +)
	60	76	80	54	68	75 (+ +)	74 (+ +)	64 (+ +)	64 (+ +)
	90	80	82	50	60	76 (+ +)	74 (+ +)	60 (+ +)	57 (+ +)
	120	72	80	52	50	74 (+ +)	73 (+ +)	56 (+ +)	53 (+ +)
RUDL	30	75	63	47	52	80 (+ +)	76 (+ +)	72 (+ +)	71 (+ +)
	60	73	64	32	47	76 (+ +)	73 (+ +)	64 (+ +)	63 (+ +)
	90	68	57	40	45	73 (+ +)	69 (+ +)	62 (+ +)	61 (+ +)
	120	71	66	41	42	74 (+ +)	70 (+ +)	59 (+ +)	57 (+ +)
RUDN	30	78	85	87	70	89 (0 0)	80 (0 0)	87 (+ +)	89 (+ +)
	60	52	86	57	71	38 (0 +)	78 (0 0)	69 (+ +)	71 (+ +)
	90	38	83	54	58	30 (0 +)	79 (0 0)	65 (+ +)	63 (+ +)
	120	31	84	57	54	27 (0 +)	80 (0 0)	64 (+ +)	59 (+ +)

Table 4: Intersection ratios of confidence intervals (intersect) of the estimate of the scale parameter and the shape parameter for Pareto thresholds of 625, 900, 1225 and 1600 points, for lead times of 30, 60, 90 and 120 minutes, for the warm and cold periods of 2017-2018. Values less than 50 are marked in red. Cases of insufficient number of situations in observations are marked with an asterisk.

Particular attention is paid to the shape parameter, the positivity of which (Pareto-type dis-

tribution of extrema) indicates the presence of a heavy tail in the distribution: the larger the value of the shape parameter, the heavier the tail and the more problematic the existence of distribution moments. It is shown that with increasing threshold, the shape parameter tends to change sign from positive to zero and, in rare cases, to negative. The zero sign in observations and forecasts at a threshold of 625 points was observed for only one radar (RUDN) during the warm period. Negative estimates of the shape parameter are even rarer; at the threshold of 625 points, such cases are completely absent.

Assuming the IR of 50% or more as an acceptable error, two **conclusions** can be drawn. First, the precipitation nowcasting system better predicts objects of extreme size in the cold season. The number of pairs (++) in the warm period according to the table 2 is about half of the cases, and in the cold - about 75%. Second, the precipitation nowcasting system most accurately reproduces the Pareto distribution of precipitation areas in the coverage areas of the RAKU (Kursk), RAVO (Voeykovo), RUDB (Bryansk), RUDL (Smolensk) radars in the warm period, and in the coverage areas to the east of the RATL (Tula) and RUDN (Nizhny Novgorod) radars in the cold period.

The last conclusion from the work can be attributed to the methodology: the extreme value theory is applicable to such objects of analysis and short-term forecasting as significant contiguous precipitation areas only with a clear understanding of the theoretical prerequisites and using suitable statistical methods and reliable data processing tools. Otherwise, the results obtained may be useless, accidental, or even harmful.

References

1. Cramer H., Mathematical Methods of Statistics. Princeton University Press (March 23, 1999), 575 pages
2. Seed A. W. A dynamic and spatial scaling approach to advection forecasting // J. Appl. Met., 2003. Vol. 42. P. 381-388.
3. Seed A.W. Modelling and forecasting rainfall in space and time // Scales in Hydrology and Water management (IAHS Publ. 287), 2004. P.137-152.
4. Bowler N., Pierce C., Seed A. STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP // Q. J. R. Meteorol. Soc. 2006. Vol. 132. P. 2127-2155.
5. Martins, E. S. and Stedinger, J. R. (2000). Generalized maximum likelihood extreme value quantile estimators for hydrologic data. Water Resources Research, 36 (3), 737–744.
6. Martins, E. S. and Stedinger, J. R. (2001). Generalized maximum likelihood Pareto-Poisson estimators for partial duration series. Water Resources Research, 37 (10), 2551–2557.

6 Verification Applications to HIW with a Focus on Spatial Methods

Question: Can spatial verification methods contribute to the proper evaluation of HIW phenomena and in what way?

HIW phenomena studied: intense precipitation, thunderstorm (lightning activity LPI, reflectivities).

6.1 Verification of forecasts of intense convective phenomena

Andrzej Mazur, Joanna Linkowska

Institute of Meteorology and Water Management – National Research Institute

Introduction

As it has already been said (cf. sub-task 2.1 report), every weather has its impact. In this part of the work carried out in the frame of AWARE Priority Project, all the activities focused mainly on the verification of the frequency of lightning discharges, predicted by means of various parametrizations. However, since every weather has its impact, each weather element can be treated as an impact source. It's just a question of scale and intensity. Therefore, the general results of the verification of convection indices - determining the possibility of hazardous meteorological situations - in relation to the measurements and calculations performed at aerological stations in Poland are additionally presented.

The verification method may be/could be/should be adapted (and specific) for each element. This report presents once again the basic assumptions of continuous (Mean Error, Root Mean Square Error) and discrete verification (FSS – Fraction Skill Score, SAL – Structure-Amplitude-Location, contingency tables) along with the idea of the VOD method, together with results of both the discrete verification and the continuous method, with the use of the VOD technique (cross-correlation/lagged correlation based on Vector Of Displacement).

Methods

Survey on (basic) methods applicable to the problem (bold marks jobs done/partially done) :

1. **SAL (Structure/Amplitude/Location) Verification**[††]
2. **FSS (Fraction Skill Score) verification**[‡‡]
3. **Categorical analysis (Contingency tables and predictands)**

all the above further on called as "discrete" analysis

1. **Standard evaluation at the grid scale** ("continuous" analysis)

[††]Wernli et al., 2008, SAL – a Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, Mon. Wea. Rev. 136(11), 4470–4487, https://doi.org/10.1175/2008MWR2415.1

[‡‡]Blaylock and Horel, 2020, Comparison of Lightning Forecasts from the High-Resolution Rapid Refresh Model to Geostationary Lightning Mapper Observations, Wea. Forecasting 35, 402-416

2. **Cross- (space-lag) correlation approach and verification**

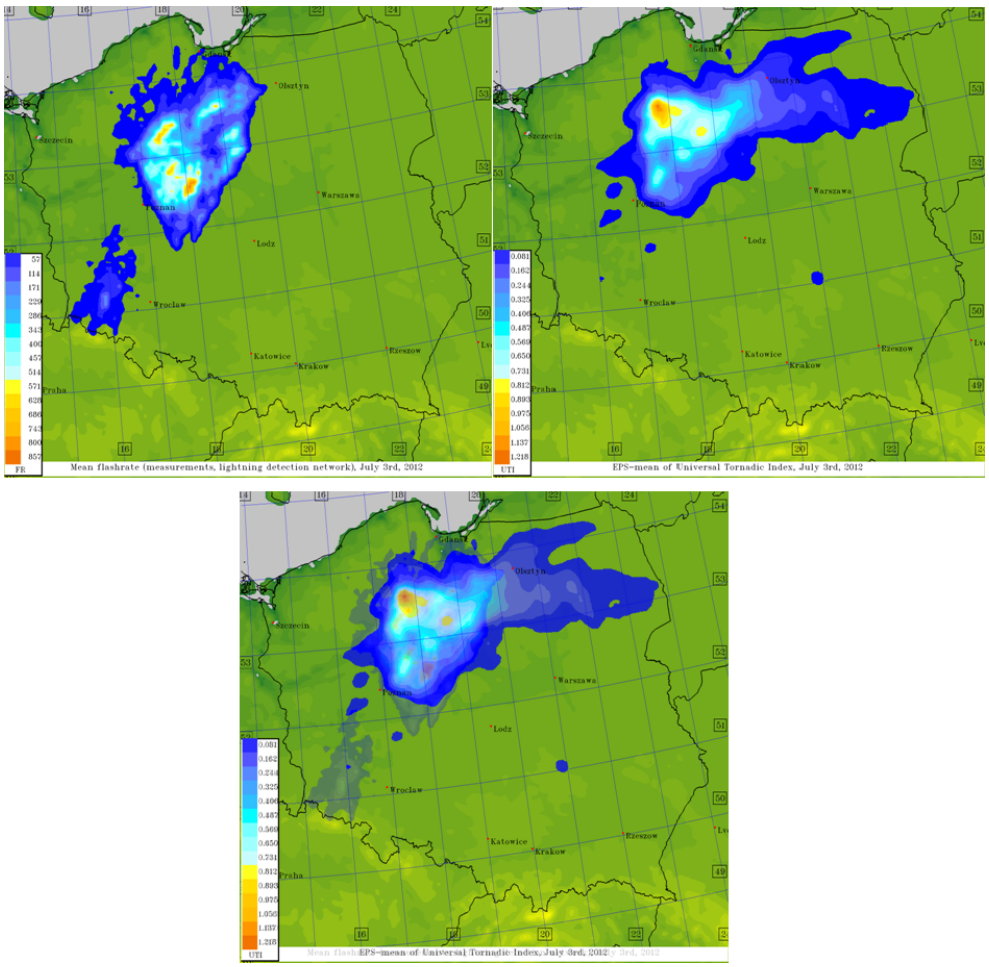


Figure 28: Basic idea of cross-correlation (lagged-correlation) approach.

When overlap the upper left (observations field) and the upper right (forecasts) panels, in most cases they do not match. It is possible to improve the forecast by using the cross-correlation (or space lag correlation) method. To do this (using the example from the figure above) one should:

1. Calculate coordinates of "centres of mass" for both distribution patterns (observations vs. forecasts).
2. Compute vector of displacement (VOD) of forecasts to observations as a difference of the two above.
3. Displace linearly every value of forecasts field by the vector of displacement.

In operational work, VOD is calculated from previous model runs (as compared to observations). It is then assumed to remain constant throughout the next run.

SAL and/or FSS and/or categorical verification for the above period has been applied (both for direct and VOD approach) to the observed and forecasted Flash Rate FR as follows:

$$FR = \left(\frac{W}{14.66}\right)^{4.54}$$

	EQS		**FAR**		**FBI**		**PFD**	
	Direct	*VOD*	*Direct*	*VOD*	*Direct*	*VOD*	*Direct*	*VOD*
2012	0.0302	0.0842	0.8832	0.8240	2.7196	2.3366	0.1736	0.1611
2013	0.0773	0.1140	0.8254	0.7920	2.4679	2.1431	0.1483	0.1232
2014	0.0299	0.0671	0.9060	0.8632	3.4946	2.6446	0.1550	0.1258
2015	0.0263	0.1022	0.8785	0.7970	2.1706	1.8439	0.1311	0.1120
2016	0.0555	0.0751	0.8532	0.8370	2.7295	2.4354	0.1592	0.1344
2017	0.0505	0.0954	0.8296	0.7976	1.9107	1.6072	0.1180	0.0978
Mean	0.0420	0.0867	0.8676	0.8221	2.3164	1.9426	0.1499	0.1283
	POD		**SUC**		**THS**		**TRS**	
	Direct	*VOD*	*Direct*	*VOD*	*Direct*	*VOD*	*Direct*	*VOD*
2012	0.2366	0.4287	0.1169	0.1760	0.0826	0.1398	0.0754	0.2551
2013	0.3245	0.4685	0.1747	0.2081	0.1249	0.1667	0.2012	0.3202
2014	0.2193	0.3863	0.0940	0.1368	0.0681	0.1096	0.0935	0.2313
2015	0.1659	0.3890	0.1215	0.2030	0.0704	0.1543	0.0538	0.2579
2016	0.2644	0.3750	0.1469	0.1630	0.1030	0.1274	0.1299	0.2157
2017	0.1981	0.3433	0.1704	0.2025	0.0925	0.1452	0.1002	0.2253
Mean	0.2349	0.3987	0.1324	0.1779	0.0898	0.1390	0.1066	0.2489

Table 5: Categorical analysis based on contingency tables.

with W being updraft velocity, calculated as

$$W = 0.3 \bullet \sqrt{2 \bullet CAPE}$$

FR is to be limited with the temperatures of top/bottom cloud temperatures, CTT and CBT, respectively.

$$if CTT > -15℃ FR = FR \bullet \left[max\left(0.01, \frac{-CTT}{15.0}\right)\right]$$

and

$$if CBT \leftarrow 5℃ FR = FR \bullet \left[max\left(0.01, \frac{15.0 + CBT}{10.0}\right)\right]$$

Another limitation is due to lack of convective clouds – if (forecasted) cloud cover is below 25%, FR is set equal to zero. Moreover, case was selected to verification if (for both observations and forecasts) maximum value over the entire domain was greater than 20 strikes/hour , and the duration of the storm was greater than 6 hours.

All the verification (both "continuous" and "discrete") was done for archive sets of observations (2011-2017).

Results are presented in the following tables and figures. Basic analysis of the results showed that VOD improved virtually all categorical predictands (like FBI, POD, THS…) from 10 up to 45%.

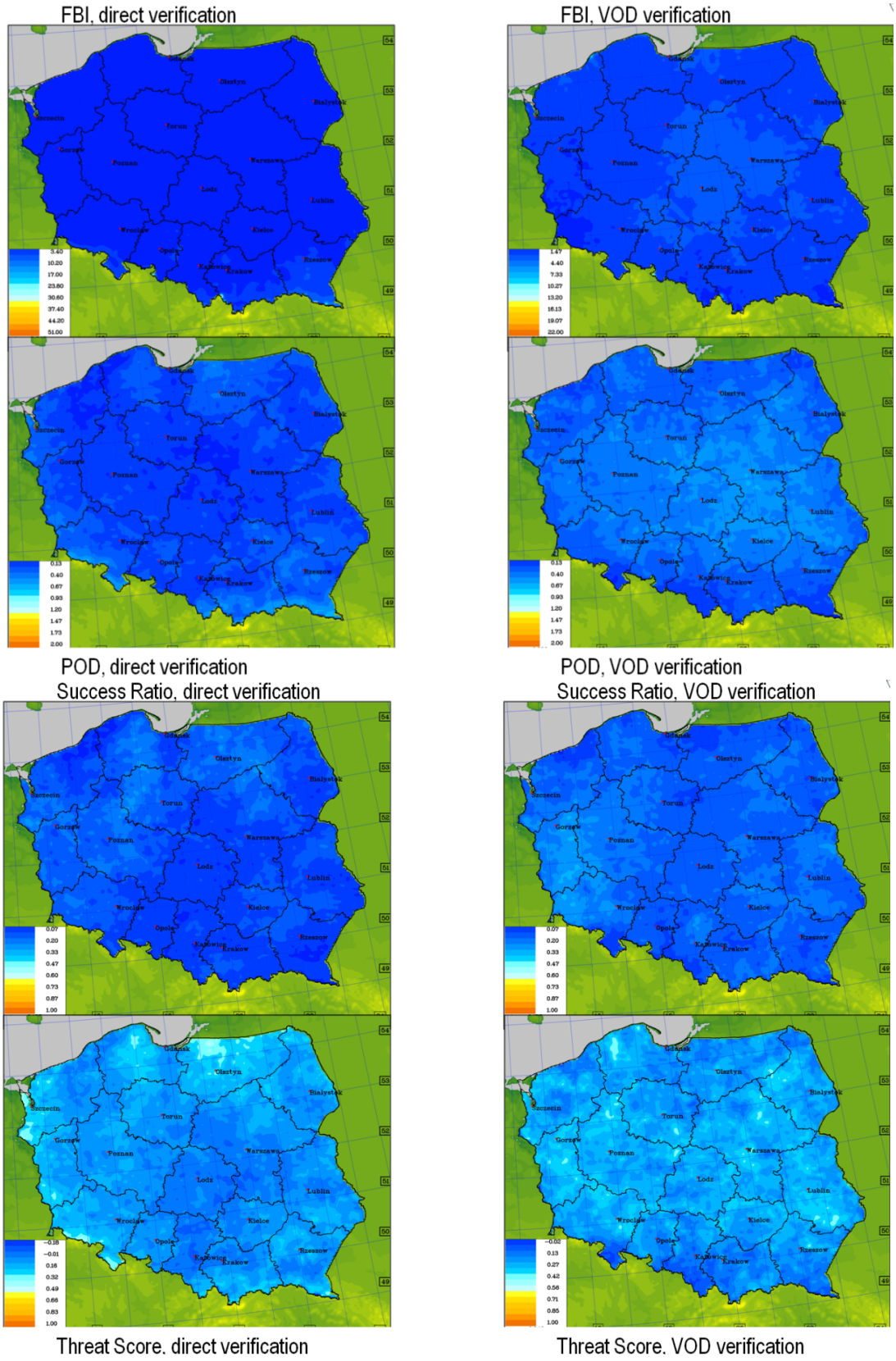


Figure 29: Examples of results of contingency tables-based verification. Top – DMO, bottom – verification with VOD applied.

The most common case is marked with bold. The parametrization of Flash Rate based on the CAPE generally overestimates FR compared to the observations. Taking into account all cases from the selected period (2011-2017), the following analysis results were obtained.

SAL with VOD applied

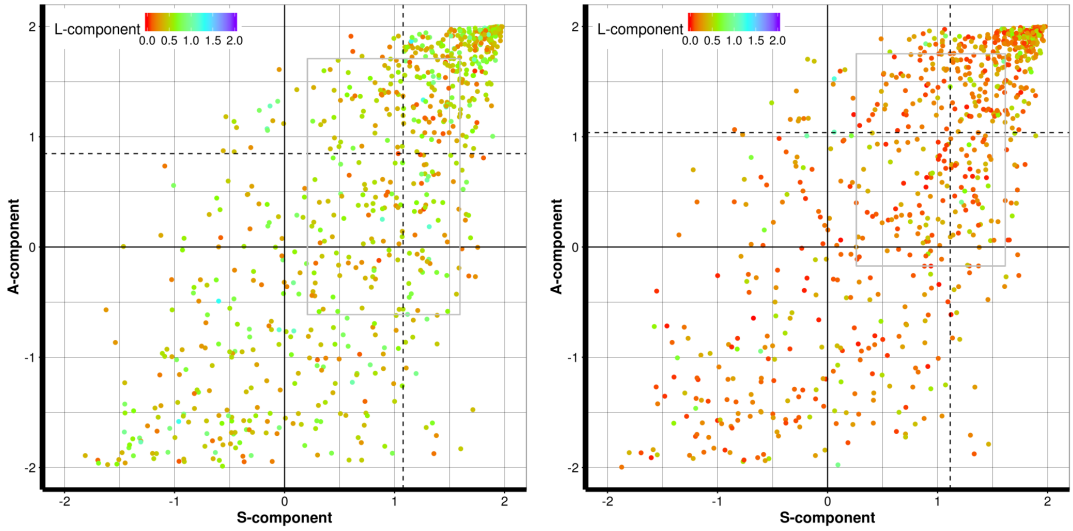


Figure 30: SAL charts for flashrate, average (2011-2017). Left diagram – direct model output results, right diagram – corrected VOD procedure.

It can be noticed that VOD forces some improvement in L-component and (to some extent) in A-component. S-component to a large extent remains unchanged. Forecasts, despite of applying VOD, are evidently overestimated. Choosing smaller domain (when SAL is to be more effective) and selection of more cases resulted, however, in no significant improvement.

Fraction Skill Scores (FSS) assessment

This method allows for direct comparison of the forecast and of observed fractional coverage of grid-box events in spatial windows of increasing size. It is supposed to be most sensitive to rare events.

Assuming probability of the occurrence of the phenomenon (in the sense of observation) as p_o, and the forecast – p_f , can be defined by the FSS according to the formula below.

$$FSS = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\left(p_f - p_o\right)^2}{\frac{1}{N}\sum_{i=1}^{N}p_f^2 + \frac{1}{N}\sum_{i=1}^{N}p_o^2}$$

Figure 31: SAL charts for flashrate, average (2011-2017). Left diagram – direct model output results, right diagram – corrected VOD procedure.

with N being number of sub-domains (or windows in overall domain).

When FSS = 0, there is no correspondence between observations and forecasts. If FSS is equal to 1, it describes a perfect match.

Again, results are shown in the following figures.

Results based on the DMO are not very good. VOD, however, significantly improves it, even up to 75%.

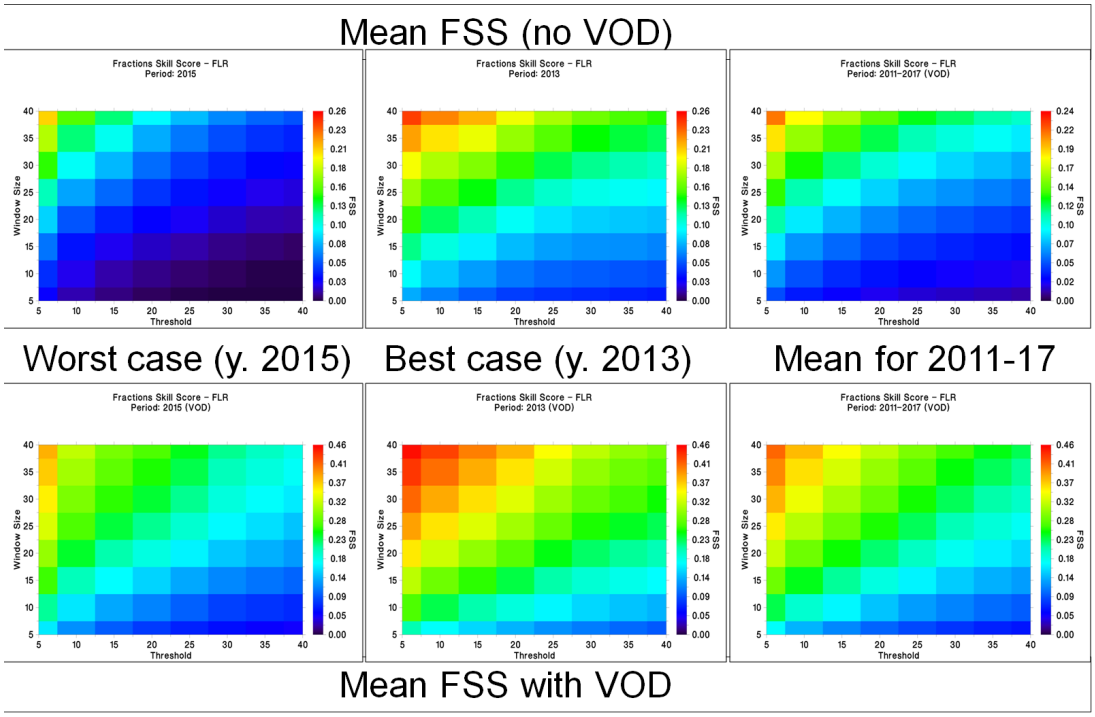


Figure 32: Values of FSS for flashrate, worst/best/average (2015, 2013, 2011-2017). Upper charts – direct model output results, lower charts – corrected VOD procedure.

	Direct			VOD		
Year	**ME**	**MAE**	**RMSE**	**ME**	**MAE**	**RMSE**
2011	2.128	**4.712**	**18.904**	1.887	**4.213**	**18.051**
2012	-2.811	**5.913**	**18.866**	-3.681	**5.027**	**17.482**
2013	-3.674	**2.184**	**10.556**	1.078	**1.949**	**9.970**
2014	-3.712	**1.516**	**9.186**	-2.192	**1.374**	**8.960**
2015	-2.023	**2.025**	**11.871**	-3.722	**1.819**	**11.391**
2016	-2.291	**3.360**	**14.695**	-0.699	**2.950**	**13.904**
2017	-1.286	**2.817**	**12.761**	-0.176	**2.015**	**11.879**
2011-2017	-1.953	**3.218**	**13.834**	-1.071	**2.764**	**13.091**

Table 6: Values of ME/MAE/RMSE for consecutive years and mean values for 2011-2017 both for "raw" (direct) values and corrected with VOD procedure.

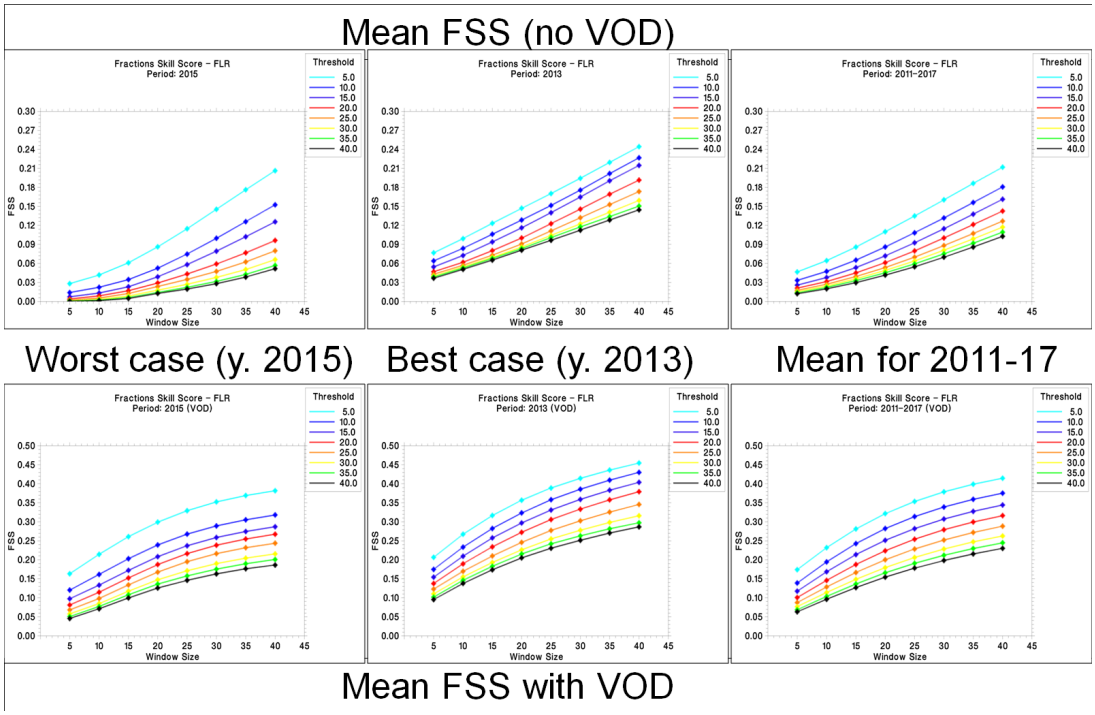


Figure 33: Values of FSS for flashrate, worst/best/average (2015, 2013, 2011-2017). Upper charts – direct model output results, lower charts – corrected VOD procedure.

Finally, "continuous" analysis requires – in general – the calculation of Mean Error (ME), Mean Absolute Error (MAE) and/or Root Mean Square Error (RMSE). Then, the basic question is - which metric is better?

RMSE has the benefit of penalizing large errors more so can be more appropriate in some cases. However, it does not describe average error alone as MAE does. Yet, distinct advantage of RMSE over MAE is that RMSE doesn't use the absolute value – which is good in many mathematical calculations. Results of calculations – both for DMO and for VOD-applied results – are presented in following tables/figures

Examples of results for year 2013, 2017 (worse, best) and means for the period are presented in following figures.

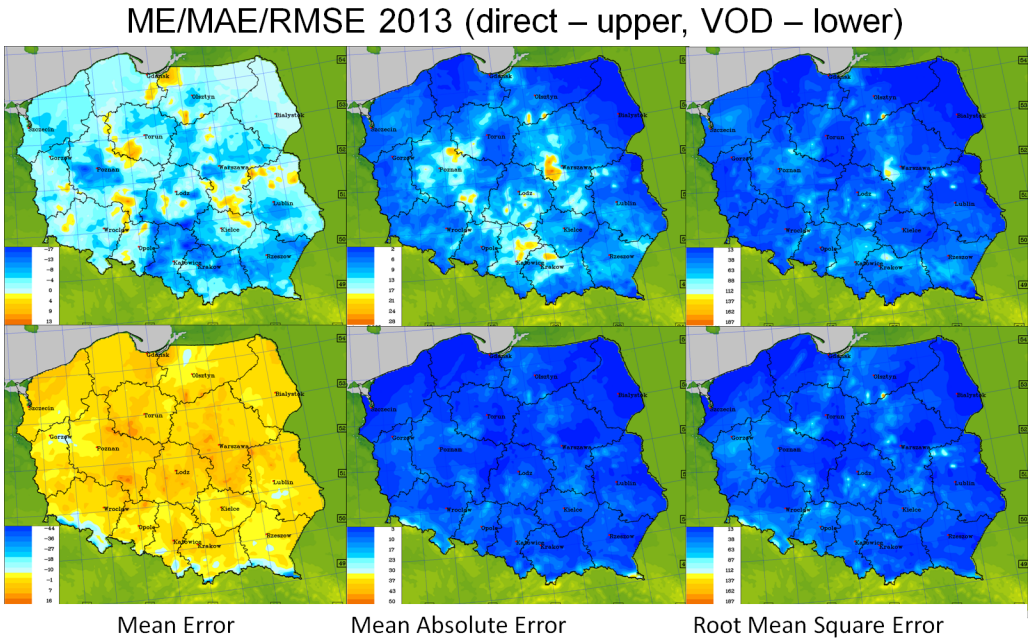


Figure 34: Values of ME/MAE/RMSE for flashrate, worst avg. year (2013). Upper charts – direct model output results, lower charts – corrected VOD procedure.

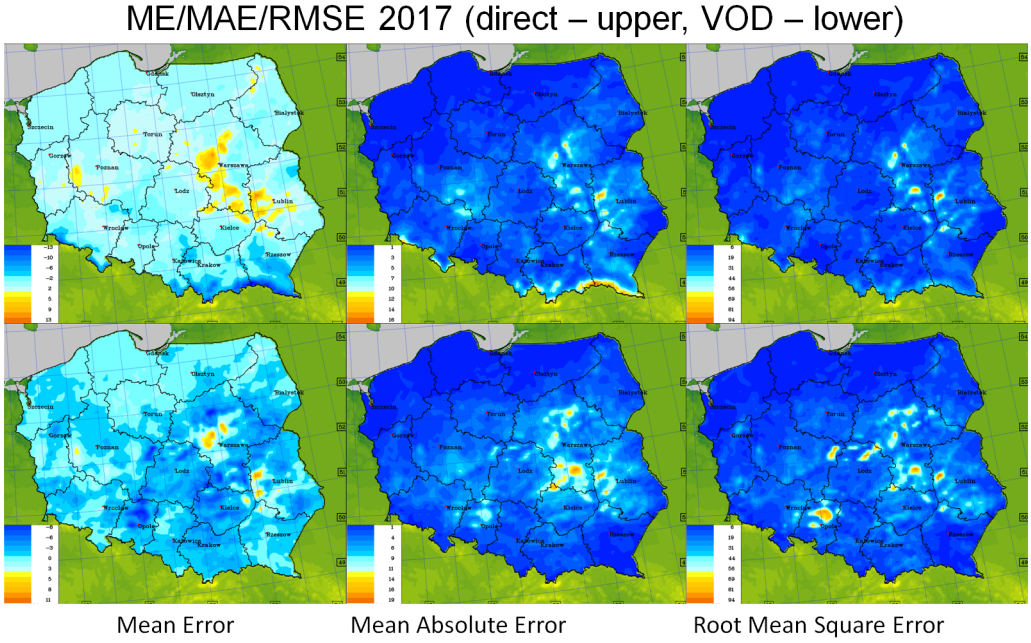


Figure 35: Values ME/MAE/RMSE for flashrate, best avg. year (2017). Upper charts – direct model output results, lower charts – corrected VOD procedure.

When consider MAE/RMSE calculated from DMO it can be seen that the worst values are apparently in mountainous regions. Maybe it is related to the fact, that it's hard(er) to predict thunderstorms in elevated terrain? When VOD procedure is applied to MAE/RMSE, slight improvement can be seen in comparison to direct verification, with a maxima of MAE/RMSE shifted towards domain centre.

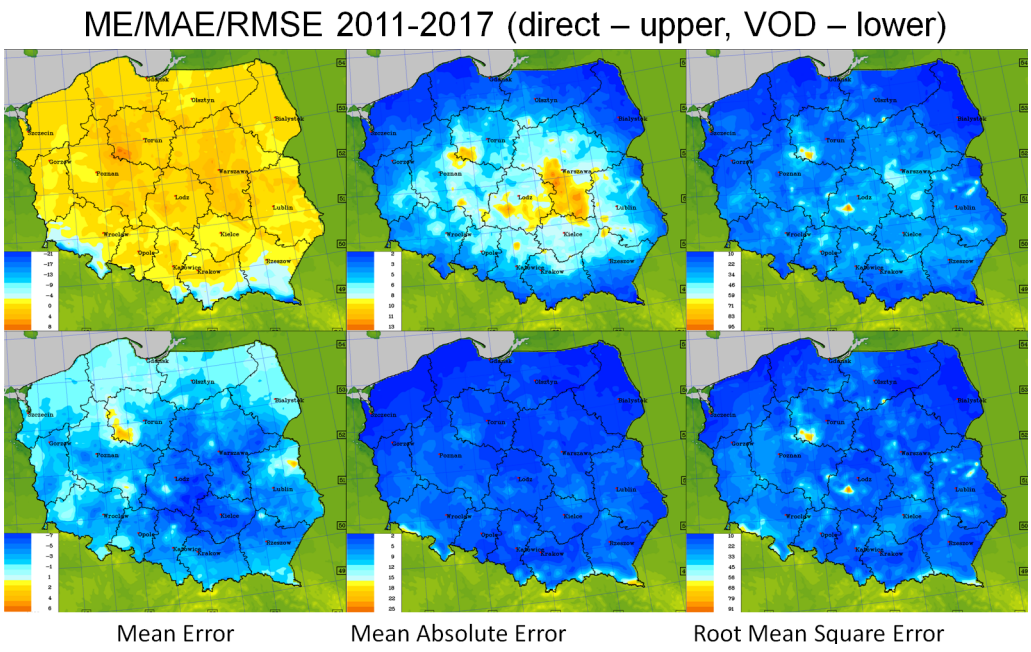


Figure 36: Average (2011-2017) values ME/MAE/RMSE for flashrate. Upper charts – direct model output results, lower charts – corrected VOD procedure.

In general VOD improves the results in all analyzed - continuous and discrete. This statement can be applied to all the cases presented.

With test period for direct- and VOD-verification extended to 2011-2017 SAL and/or FSS and/or categorical verification for the above period has been applied, both for direct and VOD approach, to four parametrizations of lightning intensity:

1. CAPE-based with cloud top/bottom temperatures correction (as described before)
2. Lightning Potential Index (LPI) (cf. U. Blahak, X.Lapillonne, D. Cattani)
3. Combination of the two above (cf. P. Lopez, D. Cattani)
4. Graupel flux at -15°C level/total ice mass (cf. J. Wilkinson, McCaul *et al.* 2019).

These four parameterizations were tested and verified against observations for two periods:

1. Case study – August 11^{th}, 2017
2. Longer period verification (June-August 2020; 7- and 2.8km only)

Results of the studies are shown in the following figures/tables.

Resolution	7.0				2.8				0.7			
Parametrization	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
ME	3.1	3.1	2.0	3.7	1.2	0.4	-0.4	0.6	-0.6	0.2	-1.5	-0.9
STD	15.7	17.7	19.9	18.0	1.9	7.4	10.0	8.2	4.0	3.4	7.2	5.7
MAE	5.7	6.3	7.0	6.4	2.3	2.6	3.4	2.8	1.3	1.0	2.3	1.8

Table 7: Continuous verification results. ME (Mean Error), STD (Standard Deviation), MAE (Mean Absolute Error).

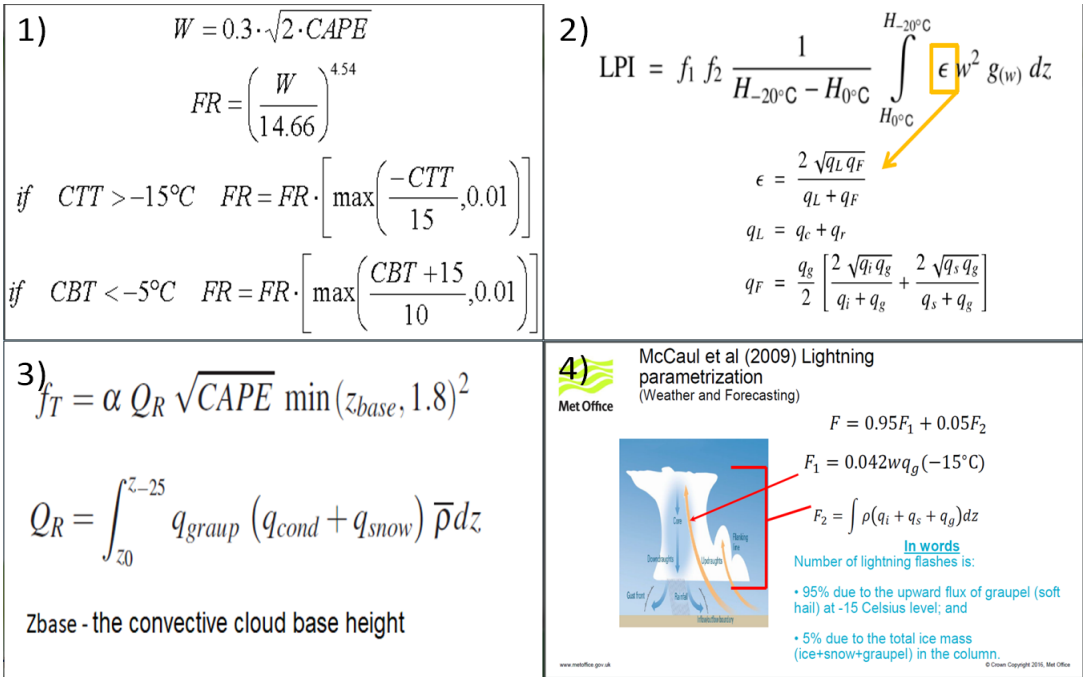


Figure 37: Basic assumptions of different parametrization of lightning intensity.

Case study August 11th, 2021

Discrete verification (Fraction Skill Score)

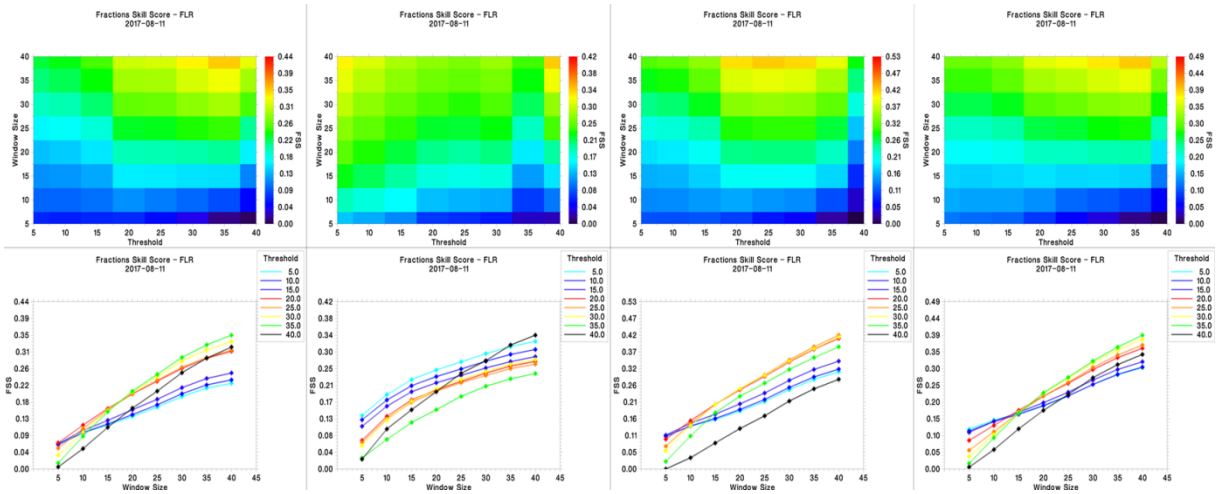


Figure 38: FSS, 7km. DMO, left to right: parametrization #1-4 (2017.08.11).

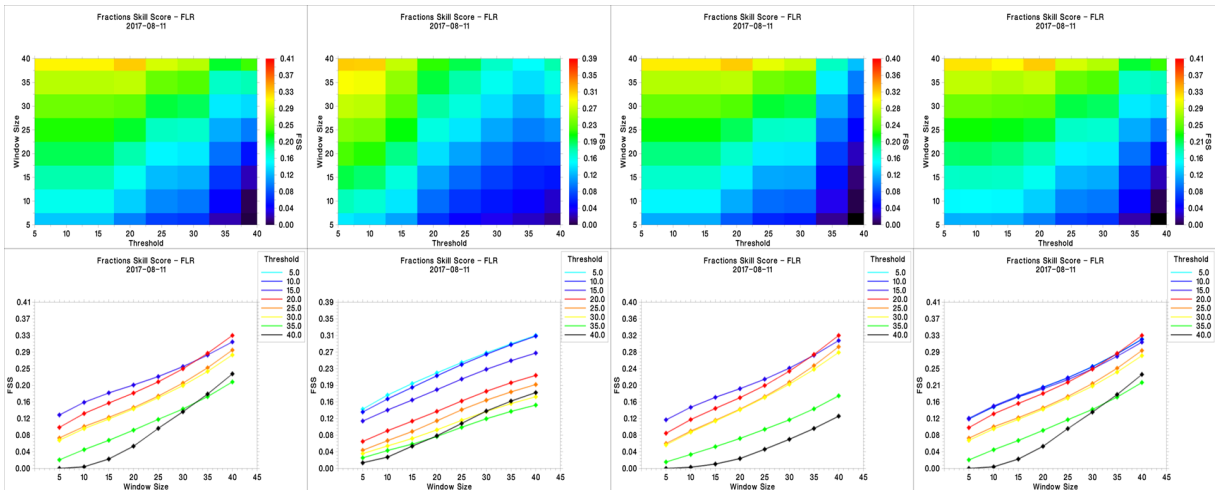


Figure 39: FSS, 2.8km. DMO, left to right: parametrization #1-4 (2017.08.11).

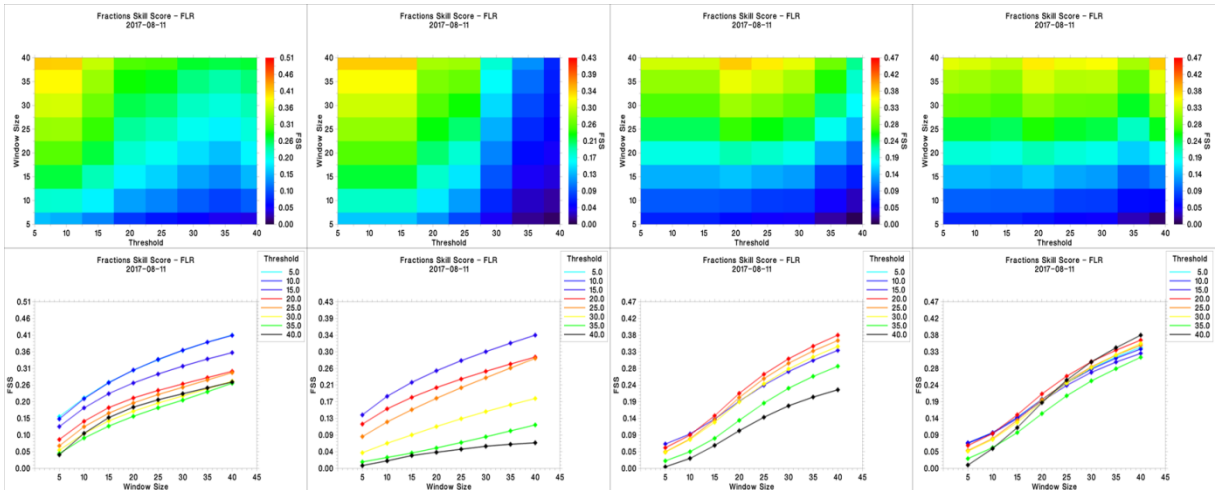


Figure 40: FSS, 0.7km. DMO, left to right: parametrization #1-4 (2017.08.11).

Summer (June-August) 2020 verification

Parametrization	EQS	FAR	FBI	PFD	POD	SUC	THS
#1	0.051	0.830	1.911	0.118	0.198	0.170	0.093
#2	0.056	0.853	2.730	0.159	0.264	0.147	0.103
#3	0.030	0.906	3.495	0.155	0.219	0.094	0.068
#4	0.030	0.883	2.720	0.174	0.237	0.117	0.083

Table 8: Verification based on contingency tables, 7km resolution.

Parametrization	EQS	FAR	FBI	PFD	POD	SUC	THS
#1	0.084	0.823	2.337	0.126	0.386	0.176	0.140
#2	0.095	0.798	1.607	0.098	0.343	0.203	0.145
#3	0.075	0.837	2.435	0.161	0.429	0.163	0.127
#4	0.067	0.863	2.645	0.134	0.375	0.137	0.110

Table 9: Verification based on contingency tables, 2.8km resolution.

Resolution	7km			2.8km		
Parametrization	ME	MAE	STD	ME	MAE	STD
#1	-9.32	3.61	23.51	-1.92	5.12	25.75
#2	-5.61	5.39	27.98	-5.66	3.49	21.41
#3	5.36	13.83	49.71	-9.23	12.16	45.61
#4	-7.28	11.87	48.63	4.18	10.71	46.83

Table 10: Continuous verification results. ME – Mean Error, MAE – Mean Absolute Error, STD – Standard Deviation.

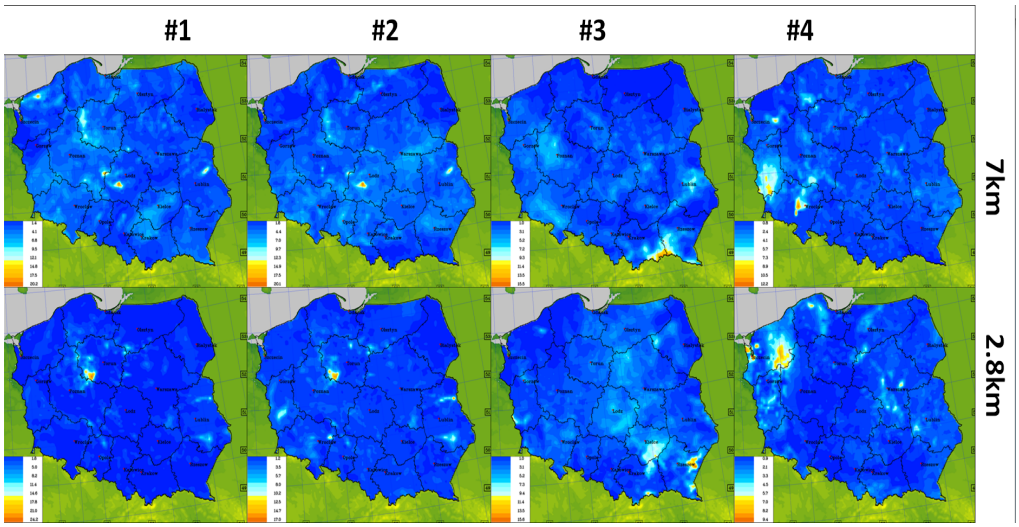


Figure 41: Flashrate continuous verification – parameterizations (summer 2020). Average values of Standard Deviation.

Considering the four compared parameterizations for continuous verification, the first two seemed to work better than others. Namely, CAPE-based parametrization worked better in coarse resolution while LPI-based – in high resolution. As far as the discrete verification is concerned (FSS, contingency tables analysis), in low resolution, results of 3^{rd} parametrization seemed to be slightly better than the other two. In high resolution first parametrization worked best. For a longer period, CAPE-based parametrization again worked better than others, while in low resolution – the one based on LPI.

Stability Indices

As far as these variables are concerned it has to be remembered that compared to the standard predicted values in the models (e.g. temperature, wind or precipitation), the possibilities of verification are significantly limited to data from atmospheric soundings and – to some extension – from satellite scans.

Atmospheric sounding (aerological) stations are located over Europe in much more scattered manner than – for example – SYNOP ones. Figure 11 presents basic available aerological stations in Europe.

Figure 42: Aerological stations in Europe used to verify forecasts of stability indices (from http://weather.uwyo.edu/upperair/sounding.html, access: September/October, 2021).

The following table presents an exemplary output from sounding at Wrocław aerological and SYNOP station, July 1^{st} , 2021, 1200 UTC

On the other hand, some stability indices can be assessed using satellite images. For instance, Showalter index is a measure of thunderstorm potential and severity. In other words, it gives a good indication where the atmosphere is unstable and where convective development may be expected. Fields of Showalter index (obtained via model forecasts, esp. in high resolution) may be compared with Meteosat 8 IR 10.8 satellite images. In some cases the discrepancy between the values of stability indices and the real situation (satellite image) can be noticed. Similarly, CAPE (Convective Available Potential Energy) – as a measure of the amount of energy available for convection – may be compared with Meteosat 8 IR 10.8 satellite images. It should be remembered that CAPE represents potential energy, and will only be used should a parcel be lifted to the level of free convection. The derived stability indices such as convective available potential energy (CAPE), lifted index (LI), total totals (TT), Showalter index (SI), and the K-index (KI) are computed from the retrieved atmospheric moisture and temperature profiles. These indices aid forecasters in nowcasting severe weather by providing them with a plan view of these atmospheric stability parameters. Forecasters use this information to monitor rapid changes in atmospheric stability over time at various geographic locations, thus improving their situational awareness in pre-convective environments for potential watch/warning scenarios.

Of course, the limitations of satellite soundings (e.g. problems with scanning in cloudy conditions, space resolution etc.) set the limits for possibility of verification of indices. Hence, it is sometimes difficult to satisfactorily define the quality of the forecast of indicators – and the possibility of the severe weather phenomenon occurring – over a large area and / or in high spatial resolution.

In this part of the report authors decided to focus on the soundings-derived values of indices for summer period (June-August) of 2020. In order to maintain a consistent image for 2.8 and 7 km resolution, eight aerological stations, located in the domain for high resolution, were selected, as listed in Table 12.

Station information and sounding indices
Station number: 12425
Observation time: 210107/1200
Station latitude: 51.13
Station longitude: 16.98
Station elevation: 116.0
Showalter index: 8.79
Lifted index: 8.76
LIFT computed using virtual temperature: 8.79
SWEAT index: 50.51
K index: 13.10
Cross totals index: 23.10
Vertical totals index: 23.60
Totals totals index: 46.70
Convective Available Potential Energy: 4.69
CAPE using virtual temperature: 5.21
Convective Inhibition: 0.00
CINS using virtual temperature: 0.00
Equilibrum Level: 864.74
Equilibrum Level using virtual temperature: 864.36
Level of Free Convection: 942.45
LFCT using virtual temperature: 943.83
Bulk Richardson Number: 1.78
Bulk Richardson Number using CAPV: 1.98
Temp [K] of the Lifted Condensation Level: 270.88
Pres [hPa] of the Lifted Condensation Level: 952.54
Equivalent potential temp [K] of the LCL: 284.16
Mean mixed layer potential temperature: 274.69
Mean mixed layer mixing ratio: 3.42
1000 hPa to 500 hPa thickness: 5235.00
Precipitable water [mm] for entire sounding: 9.65

Table 11: Output results from sounding at Wrocław (#12425), July 1st , 2021, 1200 UTC. http://weather.uwyo.edu/upperair/sounding.html, access: September/October, 2021.

Name	WMO Number	Country	Longitude	Latitude
Łeba	1210	Poland	17.50	54.75
Wrocław	12425	Poland	16.98	51.13
Legionowo	12374	Poland	20.93	52.38
Praha	11520	Czech Republic	14.46	50.00
Prostejov	11747	Czech Republic	17.09	49.46
Poprad	11952	Slovakia	20.26	49.05
Greifswald	10184	Germany	13.39	54.09
Lindenberg	10393	Germany	9.89	47.61

Table 12: Aerological stations selected for verification of stability indices.

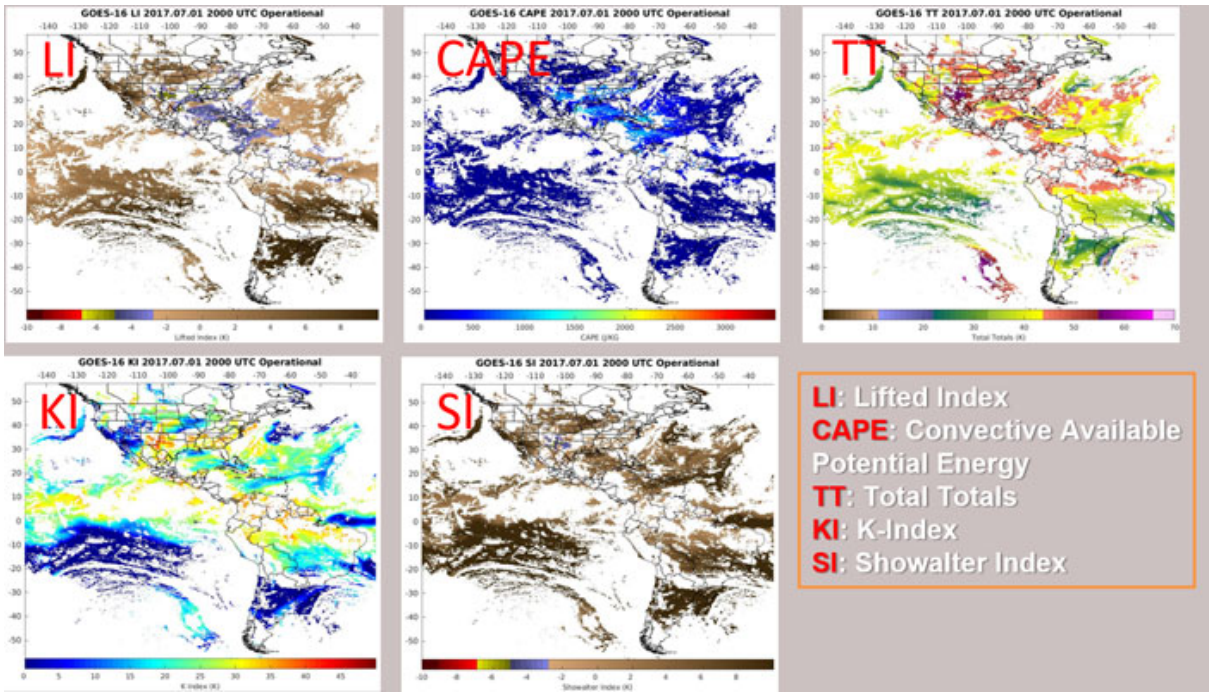


Figure 43: GOES-16 (Geostationary Operational Environmental Satellites—R Series) derived stability indices product from July 1, 2017, including lifted index (upper left), convective available potential energy (upper middle), total totals (upper right), K-index (lower left) and Showalter index (lower middle). Source: https://www.goes-r.gov/products/baseline-derived-stability-indices.html, access: October 1st, 2021.

Due to the small amount of points, lagged correlation procedure(s) has not been carried out. For the same reason, only continuous verification has been performed. For this verification, following indices have been selected: Showalter Index (SI), Lifted Index (LI), SWEAT index, K Index (KI), Totals Totals Index (TTI), Convective Available Potential Energy (CAPE) and Convective INhibition (CIN). Results are presented in the following tables.

Conclusions

Name	SI	LI	SWEAT	KI	TTI	CAPE	CIN
Łeba	-1.4	1.1	25	-8	-10.	-49	15
Wrocław	-2.0	1.3	29	6	17	38	-18
Legionowo	1.2	-0.9	19	9	11	35	-14
Praha	2.0	1.2	28	11	-18	50	22
Prostejov	2.5	1.1	-19	-10	-12	68	31
Poprad	2.1	-1.4	21	12	22	-54	-19
Greifswald	1.9	2.1	-23	12	-17	-51	21
Lindenberg	2.9	-2.5	-25	10	21	40	30
Average val.	1.2	0.3	7	5	2	10	9

Table 13: Mean error (ME) of stability indices forecasts' as compared to values at stations.

Name	SI	LI	SWEAT	KI	TTI	CAPE	CIN
Łeba	4	2	35	12	18	73	19
Wrocław	5	2	45	11	22	81	21
Legionowo	4	1.5	39	17	24	59	25
Praha	4	2	51	15	32	75	38
Prostejov	6	2	29	19	28	80	45
Poprad	4	2	40	21	35	67	32
Greifswald	5	4	39	22	29	65	34
Lindenberg	6	5	42	17	40	81	51
Average val.	5	3	40	17	29	73	33

Table 14: Mean absolute error (ME) of stability indices' forecasts as compared to values at stations.

Name	SI	LI	SWEAT	KI	TTI	CAPE	CIN
Łeba	8	4	67	23	34	139	36
Wrocław	10	4	86	21	42	155	40
Legionowo	8	3	74	32	46	113	48
Praha	8	4	97	29	61	143	72
Prostejov	11	3	55	36	53	153	86
Poprad	8	4	76	40	67	128	61
Greifswald	10	8	74	42	55	124	65
Lindenberg	11	10	80	32	76	155	97
Average val.	9	5	76	32	54	139	63

Table 15: Root mean square error (RMSE) of stability indices' forecasts as compared to values at stations.

In every parameterizations, taking into account MAE/RMSE calculated from DMO it can be seen that the worst values are apparently in mountainous regions. Authors suggest that this effect may be related to the fact that it's hard(er) to predict thunderstorms in elevated terrain. Similar correlation is hard to find considering stability indices and measurements at aerological stations. This may be, in turn, caused by the small amount of verification point and their space locations.

Comparing ME/MAE/RMSE with the boundary values of individual stability indices that determine the change in the convection situation, it should be stated that – perhaps – only in the case of CAPE the compliance of the forecast with the measurements does not substantially affect the determination of this situation. In other cases, a forecast error may result in incorrect determination of the possibility (or lack thereof) of high-impact weather. An open question remains about the compatibility of measurements (and stability indices values, which, as it should be remembered, are not DMO) on aerological stations with reality.

When VOD procedure is applied to MAE/RMSE, slight improvement can be seen in comparison to direct verification, with a maxima of MAE/RMSE shifted towards domain centre. In general VOD improves the results in all analyzed - continuous and discrete. This statement can be applied to all the cases presented.

Further works are planned to improve the Flash Rate parametrization and verify the results obtained in this way, accordingly. And last but not least important conclusion that could be drawn from all the above results is that if there is a possibility it is strongly suggested do both discrete and continuous verification.

6.2 Calibration of the Lightning Potential Index (LPI) in COSMO-1E and COSMO-2E for the production of meteogram in Data4web

Benoit Pasquier, MeteoSwiss

Introduction

Since April 27 2021 Data4web 4.0 has been used in production. It has automatized the production of the lightning pictograms that where previously done by the forecaster. It uses currently a lightning density (LD) in [*F lashes/km²/h*] produced by APN from the NWP parameter *Lightning Potential Index* (LPI) [*J/kg*]. The LPI does not provide a flash number or flash density therefore its interpretation is not intuitive and of no value to the forecaster. A transformation from the LPI into LD is necessary and was conducted by Jonas Jucker on COSMO-1 during a Master thesis in 2018 and applied to COSMO-1E and -2E. Later on the quality of the forecasts for the productions of lighting meteograms using the LD were investigated by Nicolas Schwaller for COSMO-2E and IFS-ENS. He found that the prediction with COSMO-2E were of good quality during the summer season, but the LD resulted in a strong overforecasting in Fall. To correct this overprediction he concluded that a seasonal calibration of the LPI into LD was necessary. This work is therefore exploring the different options that are possible to solve the overforecasting problem in fall.

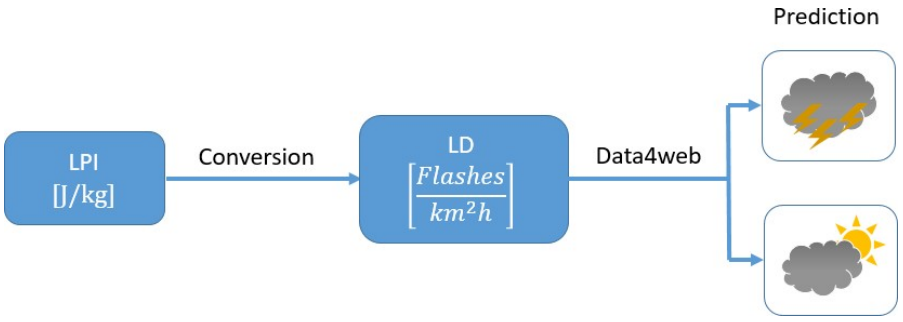


Figure 44: Different steps in the production of meteograms. The LPI is produced by NWP then it is converted in a lightning density that is then used by data4web to produce meteograms.

Lightning Potential Index

LPI [*J/kg*] is the kinetic energy of the updraft in the developing thundercloud scaled by the potential for charge separation based on ratios of ice and liquid water within the main charging zone of the cloud (Yair et al., 2010). Additional filters are necessary to remove spurious signal, they are taken without modification from Blahak (2015). As a reminder the LPI function is given below.

$$LPI = f_1 f_2 \frac{1}{H_{-20°C} - H_{0°C}} \int_{H_{0°C}}^{H_{-20°C}} \epsilon \, \omega^2 \, g(\omega) \, dz$$

Figure 45: Different steps in the production of meteograms. The LPI is produced by NWP then it is converted in a lightning density that is then used by Data4web to produce meteograms.

ω the cloud updraft

1. ϵ: the scaling factor for the cloud updraft attaining a maximum value when the mixing ratios of the ice species and the supercooled water are equal (Yair et al., 2010).

2. f_1: Neighbourhood updraft based filter, f_1 /= 0 if the majority of the neighbour grid cells ($\sim 10 \times 10 km^2$) have a maximal updraft velocity exceeding $1.1 ms^{-1}$ This value is resolution dependant but taken as it is from COSMO-DE at 2.8km gridcells

3. f_2: A neighbourhood column based stability filter for filtering the LPI with regard to graupel formation regions of intense orographic wave related clouds, where lightning activity does not develop. f_2 /= 0 if the neighbour ($\sim 20 \times 20 km^2$) grid cells all have an average of the vertically integrated buoyancy of more than —$1500 Jkg^{-1}$. Details from the comments in the COSMO-2E code: *"The buoyancy is used instead of CAPE, because CAPE is by definition 0 in the column of an upright convective updraft core and only counts the positive buoyancy parts of a parcel ascent. It is only a measure of instability, not stability. Buoyancy also takes into account the negative buoyancy contributions and can better distinguish between a convective updraft (buo = $\approx$ 0 J/kg) and orographic clouds in stable stratification (buo << 0 J/kg).* ***A threshold of -1500 J/kg has been determined by U. Blahak based on COSMO-DE summer season data of 2014.*** *If the smoothed buo falls below this threshold, LPI is set to 0."*

4. $g(\omega)$: velocity based filter function within the column $g(\omega)$ /= 0 if $\omega > 0.5 ms$—1

Filters in the LPI formula

It is important to notice that the filters were not adapted from the work of Blahak (2015) that was done on COSMO-DE that has a resolution of 2.8km between 28.7-16.8.2014 for f_1 and 6.10-23.10.2014 for f_2 and $g(\omega)$, with two runs a day at 00UTC and 12 UTC and a leadtime of 11h. The LPI was compared to the time averaged observed flash rate +-15 min around the date in the unit $1/(km^2 min)$. They were not adapted for the COSMO-1 or COSMO-2E resolution. Additionally the timeframe that was used for the determination of the thresholds of the filter was short and done 'by eye'. The solution implemented by APN was to set a threshold in [J/kg] on the values rather than adapting the filters.

COSMO-1 Lightning Potential Index to Lightning Density conversion

The first study on the transformation from LPI to LD was done in 2018 by Jonas Jucker in a master thesis, where more details can be found on the method.

Model

COSMO-1 was used to calibrate the transformation from LPI to LD. Reforecast on the month of July 2017 with one run per day initialized at 00 UTC with a leadtime of 24h were used. Instantaneous values of LPI were calculated at every exact hour for each gridpoint. The LPI was then upscaled from $1.1 \times 1.1 km^2$ to $6.6 \times 6.6 km^2$ by taking the maximal value. Finally different threshold were used between 0 and 20 J/kg, when the gridpoint LPI value would be above this threshold. It would be considered that there is at least a flash on the gridpoint per hour.

Observations

Jonas Jucker used lightning data from the European Cooperation for Lightning Detection (EUCLID) (Schulz et al., 2016). It detects 95% of all Cloud to Ground (CG) and 45 % of Cloud to cloud lightning strikes with an location accuracy of 500m. The network provides

data for central Europe, including Switzerland. The observation were first gridded on the $1.1km^2$ grid and then summed on a $6.6km^2$

Results

To evaluate the quality of LPI in COSMO-1 the proposed verification procedure by Wilkinson (2017) was applied. The predictions are compared to the observations using a contingency table for each threshold. A coverage score (SEDS) that measures the skill of the forecast in covering the correct number of gridpoints with lightning activity was used. It is explained in section 5.1

As stated before, in the definition of the LPI the filters (especially the updraft velocities) were defined for the COSMO-DE model (Blahak 2015) that has a larger resolution than COSMO-1. This results in an under pruning of predicted flashes. Assuming that a value of LPI > 0 for a gridpoint would be equal to at least one flash is a large overprediction. To correct this bias without adapting the filter to the domain resolution of the model (out of the scope of his work and this one as well), Jonas Jucker used the value of the LPI for the maximal coverage score as a threshold under which all values of LD would be set to zero.

He also did a categorization of the different lightning intensities, but this is not in the scope of the production of Flashes/no Flashes meteograms. Additionally the intensities scores were poor with LPI values calculated only every hour. This means that until the capacities to calculate the LPI more than every hour are created, the intensities will be unreliable. A good forecast in terms of intensity is only possible when capturing the right coverage as well. This means that having a right coverage is the first step towards a further usage of the promising LPI.

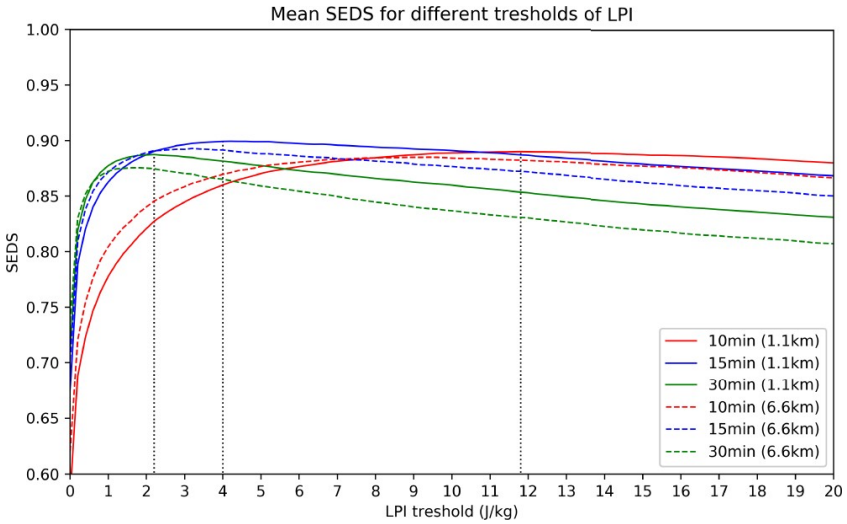


Figure 46: Mean SEDS for thresholds of LPI from 0 to 20 J/kg. The vertical grey dotted lines mark the maximum SEDS for 1.1 km resolution. The green continuous line is what interest us. Its maximum value is for a threshold just below 2 J/kg (Jonas Jucker Master thesis).

LPI [J/kg]	LD [Flash/km2/h]
<2	0
]2-7]	1
]7-14]	2
]14-22]	3
]22-40]	>3
]40-85]	>5
>85	>10

Figure 47: Thresholds for the 7 categories estimated by Jonas Jucker. The flash-no flashes threshold is the one interesting for the creation of meteograms.

Another problem that was found is the overforecast that takes place right after the model initialization. Jonas Jucker assumed that this was because of the latent heat nudging that takes place up to 1h and 15min leadtime. This is not a problem as the predictions at such small leadtime won't be used and proves again the necessity for a nowcasting of lightning strikes.

Evaluation of the Lightning density in COSMO-2E for the production of Meteogram by Data4web

The conversion LPI-LD calculated by Jonas Jucker on COSMO-1 was then used on COMSO-2E. Nicolas Schwaller has evaluated its use for the production of meteograms (flash vs. noflash). He has tried to find which spatial upscaling area was optimal and the probability threshold above a lightning pictogram should appear. Only the problematic results are presented here. The rest can be found in his presentation and report.

Observations

He used the observation from the Meteorage detection Network (part of EUCLID). That is able to detect 98% of CG strikes and between 30 to 50% of CC strikes with a medium accuracy of 100m. The observations were gridded to the nearest COSMO-2E gridpoint, by aggregating observations that occurred +- 30min around each exact hour.

***M*odel**

The LD in $flash/km^2/h$ was available at every exact hour its validity is for +- 30min around the reference time. Two runs per day are confronted to the observations up to a leadtime of 48 hours for the period of summer 2020.

Method

Some data post-processing was done prior to evaluation of the prediction. First the scale of reference of the prediction was change to account for the spatial uncertainty in the model prediction and compensate for the underdispersiveness of the model. A spatial square window of area A centred on each gridpoint is used, the data of each gridpoint is replaced by the sum of the number of flashes for observations and by the maximum value of LD for the model prediction, computed over all gridpoints inside the window. A change of the time scale of reference can be performed in the same way, with a 1D window in the time dimension.

Probability of having more than one flash in an area A around each gridpoint are then computed for each gridpoint as the fraction of members with a LD of at least 1 *Flash/gridpoint/h*. The upscaling produces increased and smoothed probabilities. Then a threshold is used as a separation for the flash no flash prediction.

Fall overforecasting

The quality of the prevision in summer were satisfying, but the results for fall showed a strong overforecasting of lightning activity as seen in fig 4. The observation bias in fig **??** shows also an overprediction for a 20% threshold and an upscaling of 20x20km^2 (the values that were chosen based on summer forecast for the prediction of meteograms. Changing the threshold on the number of flashes predicted per gridcell from 1 to 2 for the creation of a meteogram with a meaning: *at least one flash* is not coherent. This is why a change along season of the conversion table LPI-LD used in fig 1 is necessary.

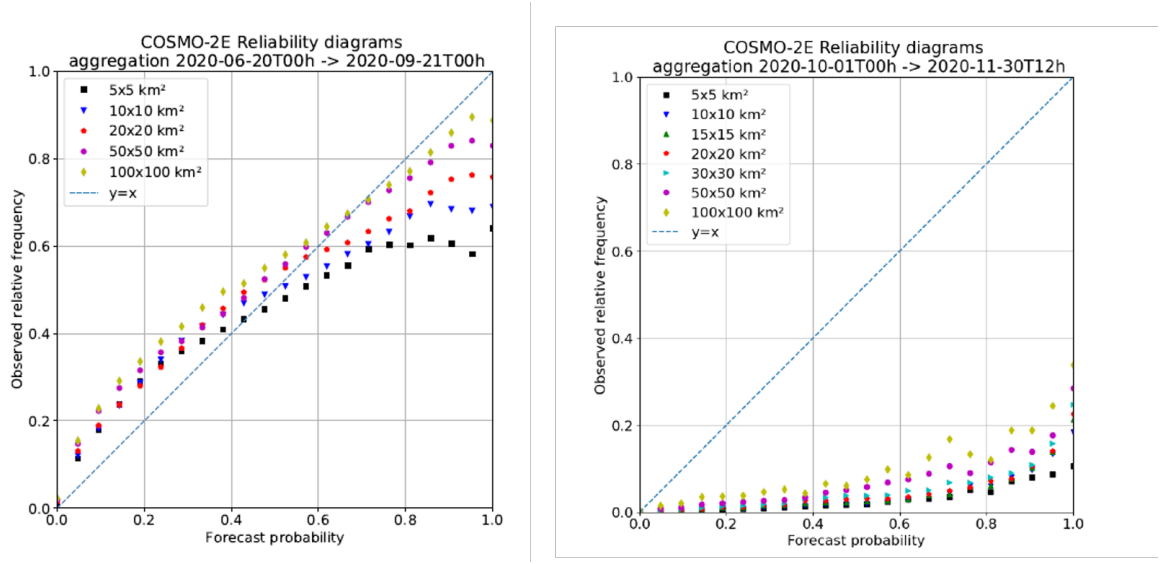


Figure 48: Reliability diagrams for summer 2020 on the left and fall 2020 on the right. Both are for a detection level of 1 *flash/km^2/h*. The situation in Fall shows a strong over forecasting.

Monthly calibration

The same methodology that Jonas Jucker used for the definition of the flash/noflash threshold is used on a monthly period to see if there is an evolution in the threshold that would maximize the SEDS along the different months.

Results

Forecasts are then compared to the observations using the Symmetric Extreme Dependency Score (SEDS). It compares the skill of the forecast to a hedged forecast, which is a forecast that predicts lightning for every gridcell in the model domain. SEDS varies between 0 for a hedged forecast and 1 for a perfect forecast.

Additionally, edge cases where either p = 0 or q = 0 are considered as zero skill and SEDS is set to zero. This would be situation where there is no flashes observed but some are predicted, or the opposite, flashes are predicted but none is observed. The cases where no flashes is observed or predicted that means there is only true negatives are removed from

the calculation of the SEDS as they would affect greatly the average SEDS if the SEDS would be set to one in this case.

For the interpretation of the results one must be careful of some characteristics of the SEDS. It is important to note that SEDS is dependant on the Base Rate and will decrease with decreasing base rate and that also it varies slowly for increasing coverage bias and number of observations in the model domain (Ferro 2011). Also, the probability that a random forecasting system produces a contingency table very different from the expected contingency table decreases as n increases.

Doing verification one must be careful to the problem of the so-called double penalty. A predicted flash that is displaced in space or delayed in time is scored worse than either a complete miss or a false alarm since it is penalized as both at once. Also, there is almost no flashes during the winter season, therefore they are of no statistical relevance.

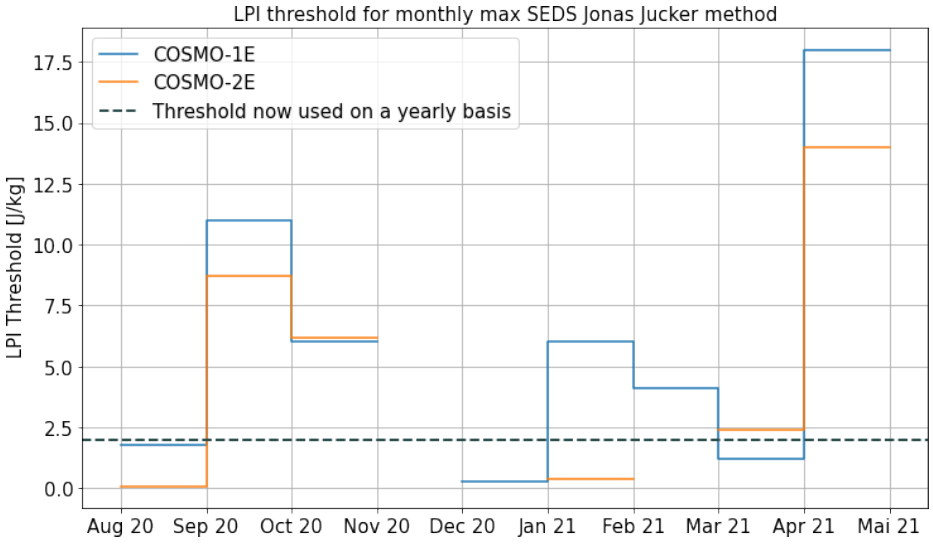


Figure 49: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E on a monthly basis using the methodology of Jonas Jucker.

Discussion

The first result is that the currently used threshold that was calculated on one summer month for COSMO-1 is not appropriate for the months of September 2020 to April 2021. For COSMO-1E the threshold is close to the current one in August, which is in concordance with the results previously obtained for a summer month and this scale of upscaling. However the best threshold for COSMO-2E in August 2020 is close to zero and this could be because the filters in the implementation of the LPI in section 2 were calculated for COMSO-DE that had a gridsize of 2.8km which is close to the 2.2 km grid of COSMO-2E. The filters should be adapted when changing the resolution of the model, but they were not adapted for COSMO-1E and this is seen as the threshold that maximize the coverage score is not zero.

The second straight forward results that can be noticed is that the value for the optimal threshold in August 2020 is not the same in fall and winter months with a big variability between every month and also some missing values, this is probably because of the sparsity of the data thus having no statistical relevance. This would mean that the LPI has been developed to predict lightning strikes well during summer months were the vast majority

of lightnings happened. But that is overpredicts lightnings during the colder season and Especially the transition months. A month to month calibration would be more appropriate to find the threshold for each period, but data is only available for a year, that is why season calibration would make more sense.

Seasonal calibration

The same methodology that Jonas Jucker used for the definition of the flash/noflash threshold is used on a monthly period to see if there is an evolution in the threshold that would maximize the SEDS along the different months with the exception that an upscaling is done to take into account the fact that even though the LPI is a localized event, it has a wide impact, as the noise of thunder can scare people in a wide area. A compromise was found for the area of the upscaling to be 7km x 7km. It was considered that the thunder is really loud in such an area.

The methodology that was used to determine the threshold that would be the best to predict at least one flash per grid point per hour was also used for two and three flashes per grid point per hour.

Model

The control run of COSMO-1E and COSMO-2E is used with runs on a respective leadtime of 33h and 48 hours, with runs respectively every 3h and 6h. The LPI is still an instantaneous value calculated at every hour for each gridpoint. It is upscaled on $6.6km^2$ which is a neighbourhood of 7 gripoints wide for COSMO-1E and 3 gridpoints wide for COSMO-2E.

Observations

The observations are from the Meteorage detection network. They are gridded to the closest gridpoint and then upscaled by summing for each gridpoints the observations on the neighbouring gridpoints on a 7kmx7km window centred on the grid point. After upscaling the observations, there is 49 times more observations than before.

Results

COSMO-1E	Threshold for flashes/km2/h		
Month	1	2	3
August 2020	2.9	4.8	11
September 2020	4.9	11	17
October 2020	6.8	9.9	18
November 2020			
December 2020	0.3	0.2	0.2
January 2021	3.1	3.1	3.9
February 2021	4.1	4.1	4.1
March 2021	42	46	1.2
April 2021	6.2	6.7	7
Mai 2021			
June 2021			
July 2021			

Figure 50: Threshold that maximizes the SEDS for COSMO-1E for one, two and three flashes per hour per km2.

COSMO-2E	Threshold for flashes/km2/h		
Month	1	2	3
August 2020	0.1	3.6	4.9
September 2020	11	11	12
October 2020	8.3	3.6	1.9
November 2020			
December 2020	4.1	0.1	0.1
January 2021	21	21	21
February 2021	0.2	0.2	0.2
March 2021	7.7	7.4	28
April 2021	33	70	70
Mai 2021			
June 2021			
July 2021			

Figure 51: Threshold that maximizes the SEDS for COSMO-2E for one, two and three flashes per hour per km2.

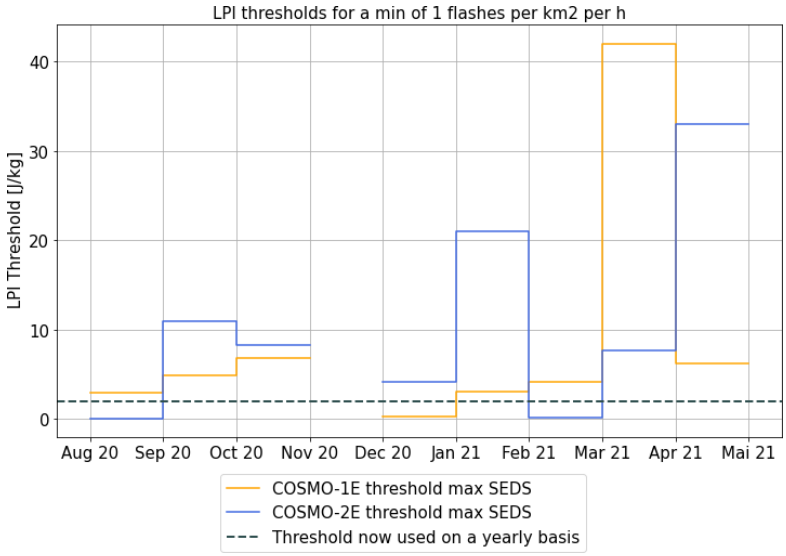


Figure 52: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E to predict at least one flash per gridpoint per hour on a monthly basis.

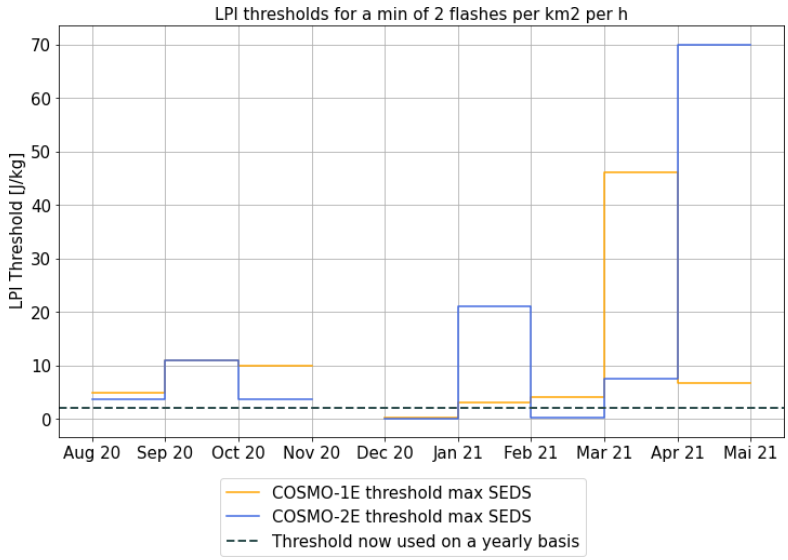


Figure 53: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E to predict at least one flash per gridpoint per hour on a monthly.

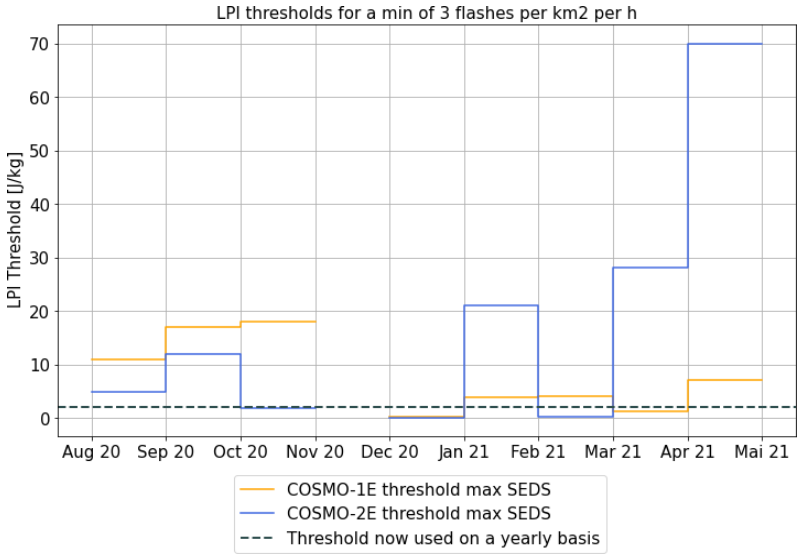


Figure 54: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E to predict at least one flash per gridpoint per hour on a monthly basis.

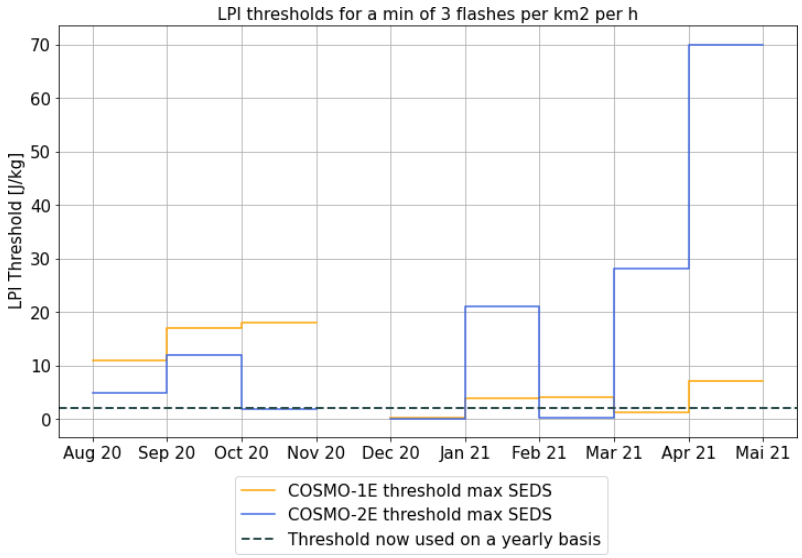


Figure 55: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E to predict at least one flash per gridpoint per hour on a monthly basis.

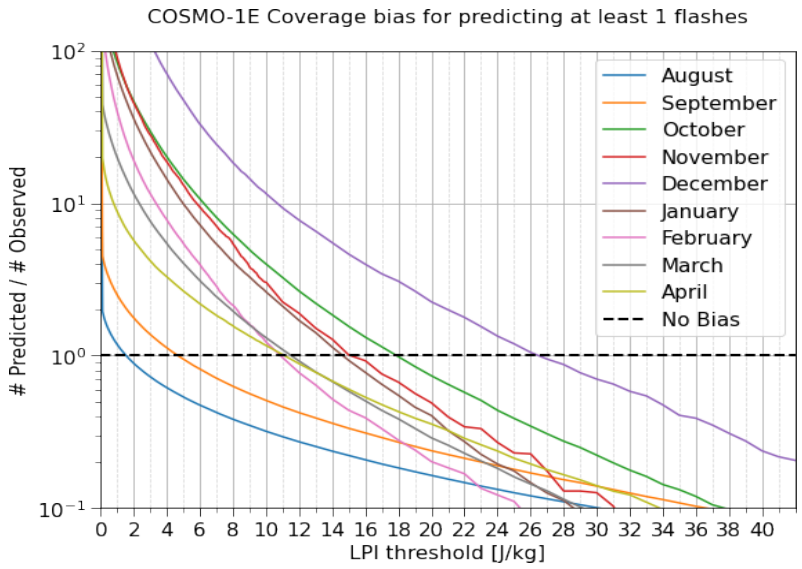


Figure 56: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E to predict at least one flash per gridpoint per hour on a monthly basis.

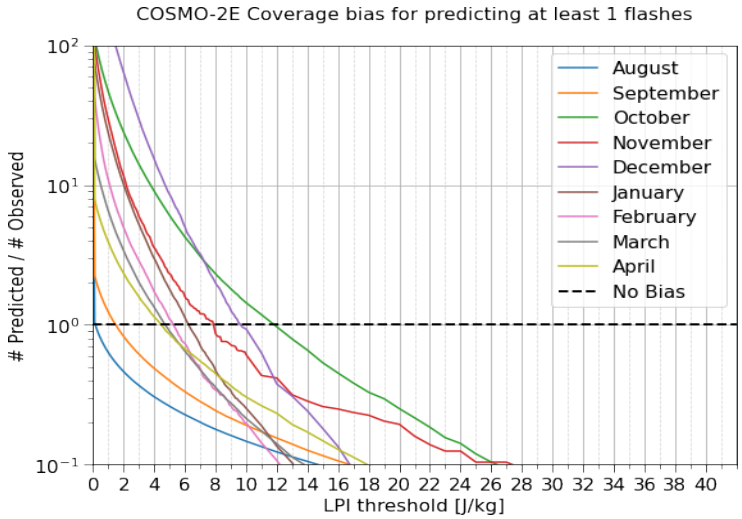


Figure 57: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E to predict at least one flash per gridpoint per hour on a monthly basis.

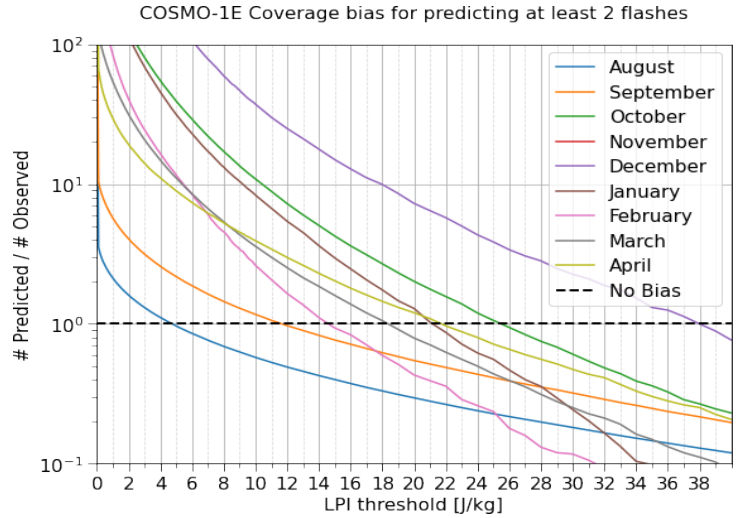


Figure 58: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E to predict at least one flash per gridpoint per hour on a monthly basis.

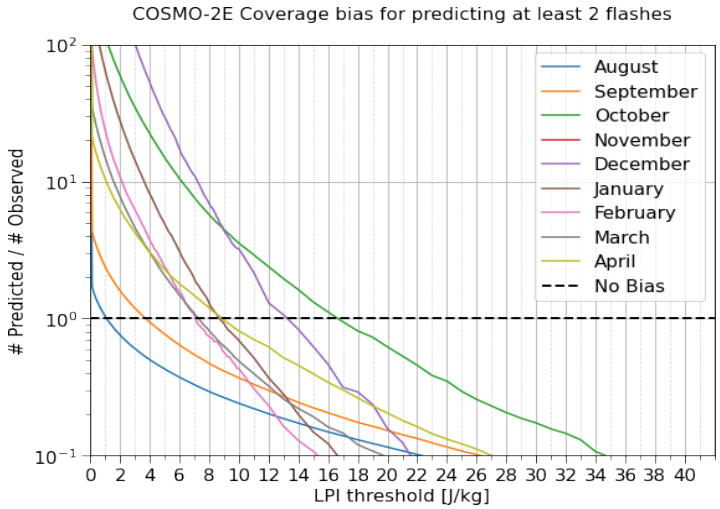


Figure 59: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E to predict at least one flash per gridpoint per hour on a monthly basis.

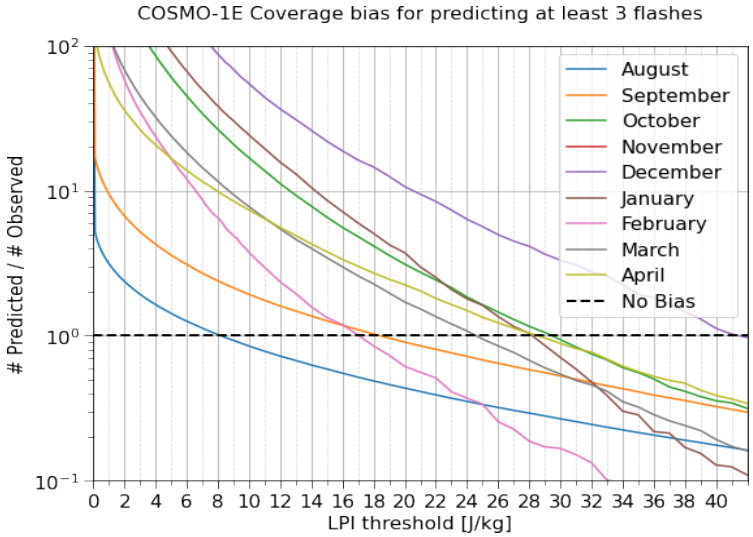


Figure 60: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E to predict at least one flash per gridpoint per hour on a monthly basis.

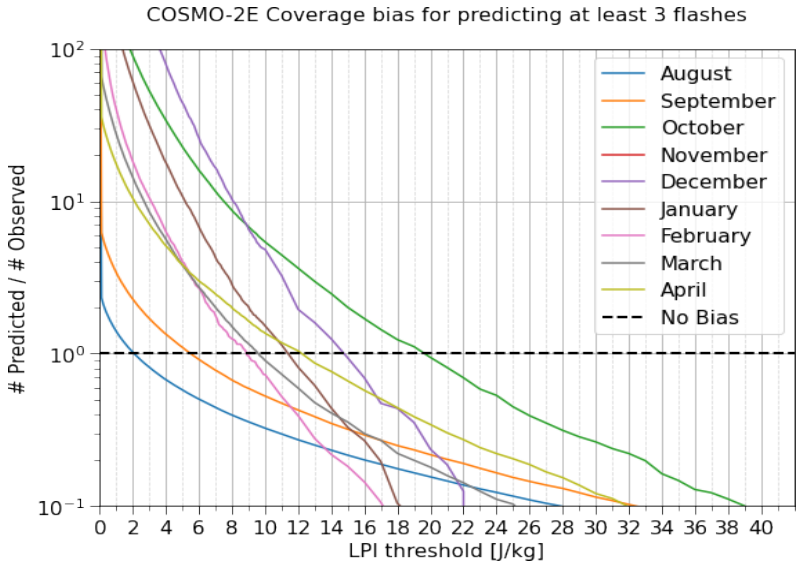


Figure 61: Threshold that maximizes the SEDS for COSMO-1E and COSMO-2E to predict at least one flash per gridpoint per hour on a monthly basis.

Discussion

It can clearly seen from the coverage bias plots and the threshold that a yearly threshold for the LPI can't be used for good prediction of thunders. Especially during the transition month of September October and Mars-April, but also during the winter season. The threshold is usually lower for COSMO-2E compared to COSMO-1E maybe because the filters that are gridsize dependant, are more efficient because they were calculated for a 2.8km gridsize, that is closer to 2.2km than 1.1km.

The coverage bias varies massively for cases showing only little lightning activity. It was already noticed by Jonas Jucker. His hypothesis was that it was because of the double penalty but using spatial upscaling and time upscaling that should reduce significantly the double penalty has only a mild effect on the threshold in the month with little lightning activity. This speaks again for the hypothesis that the filters are not adapted and do not filter some type of weather that crated spurious signal during the transition and winter month.

One could ask about the definition of the lightning meteogram, if the meaning of at least one flash per km2 per hour if meaning full and should not be set to a higher number. The interpretation of the lightning pictogram could have a tendency to be interpreted as "stormy weather". This is why maybe to create the lightning pictogram, maybe two or three flashes per hours per km2 per hour should be used as a limit.

It is important to note also the effect of the leadtime on the threshold, with increasing leadtime, the localization of the convective cell will diminish, and uncertainty will rise and thus using longer leadtimes will lead to a higher threshold to compensate for the wrong values created by the uncertainty. Thresholds should be adapted for each leadtimes as the skill of the model to predict convective cell decreases.

Setting a threshold would mean that in the type of weather that do not need to be filtered, flashes would be ignored and in the situation that would need to be filtered, wrong prediction will be made. Therefore if a good prediction and communication to the public on the thunderstorm using the LPI wants to be used, it will go through an evaluation for prob-

lematic situations in Switzerland and filters parameters to be adapted to the grid of each model.

Conclusion

The Filters in LPI formulas are more efficient in COSMO-2E than COSMO-1E, but for both models they show inefficient pruning of spurious signal especially in orographic precipitation. An adaptation of the filters threshold should be done. However, the production of meteogram can be improved immediately by changing flash threshold to suggested level along seasons. Or if the LPI would be directly available in Data4web.

Using a threshold does not rely on physical parameters and is always a trade-off. In weather type where the values are well filtered by the LPI formula setting an additional threshold would prune unnecessary values and create missed and in weather types where the filters in the LPI formula are not efficiently removing spurious signal, setting a high threshold would be good to prune false alarms. The two cases are opposing each other. Another aspect that speaks for the adaptation of the filters in the LPI formula is that the conversion from LPI to LD depends on the degree of freedom of the methodology (time area upscaling, leadtime) and is therefore arbitrary and the best calibration will depend of the methodology for the production of meteograms. Therefore, the pruning of the spurious signal should be based on physical parameters from the model and not the methodology used for the production of meteogram. A way of overcoming this problem would be to use directly LPI in Data4web instead of using the lightning density.

Future work

Some ideas for future improvements:

1. Do a full year study on the remaining month (Mai –July 21)
2. Adjust the filters or try out the new qg-filter from ICON at DWD.
3. Evaluate the suggested thresholds with probabilistic means now that data is available.
4. Use other NWP inputs or topographic descriptor in combination with LPI to produce lightning meteograms.

6.3 MODE verification of ensemble precipitation forecasts at RHM

Anastasia Bundel and Elena Astakhova (who wasn't participant of AWARE, but participated in writing this chapter), RHM

Papers published within this theme:

Bundel Anastasia, Elena Astakhova, Elizaveta Olkhovaya, Alexander Kirsanov and Dmitry Alferov, Spatial verification of a regional ensemble precipitation forecasting system at the Hydrometeorological Research Center of the Russian Federation using a free verification package, MET // 2022 *IOP Conf. Ser.: Earth Environ. Sci.* 1023 012001 **DOI** 10.1088/1755-1315/1023/1/012001

A.Yu. Bundel , A.V. Muraviev, E.D. Olkhovaya, Overview of spatial verification methods and their application to ensemble forecasting, Hydrometeorological Research and Forecasting, 2021, No. 4 (382). PP. 30-49. DOI: https://doi.org/10.37162/2618-9631-2021-4-30-49 [In Russian]

Introduction

The applicability and usefulness of spatial verification methods were tested for a limited-area ensemble prediction system ICON-Ru2-EPS. The system is based on the ICON model [8], which is a nonhydrostatic model with an icosahedral grid.

Data and verification setup

Ensemble prediction system

ICON-Ru2-EPS is a convection-permitting system with a horizontal resolution of about 2.2 km and 65 vertical levels. The integration domain covers the Central Federal District of Russia (approximately 50-60°N, 29-43°E) and contains 200416 grid cells. The grid is a rotated one and the North Pole is shifted to 35°N 215°E. Initial and lateral boundary conditions are obtained from the global ICON runs with a grid step of about 13 km and 90 vertical levels (provided by the German Meteorological Service). The lateral boundary conditions are updated every 3 hours. The verification experiment was held with a research version of the system. The ensemble comprised eleven realizations. One realization was a control one (with no perturbations included) and ten other realizations were generated only through model perturbations. To this end, the parameters of physical parameterizations of several processes like convection, turbulence, soil processes, etc. were stochastically disturbed. A set of variables to perturb was defined from model sensitivity experiments. The tuning parameters were perturbed within a meaningful range so that the forecast skill on average did not get worse. The forecasts were run for 48 hours starting from 00UTC for summer conditions (July 2021). The output was prepared on the regular latitude-longitude grid with a step of 0.02°.

Observations

The radar composite over the Central Russia was used as gridded precipitation observation dataset. The radar data were provided by the Central Aerological Observatory of Russia. The radar composite data with 1 km grid mesh were interpolated to the model grid with 2.2 grid mesh using the nearest grid point approach. Ten minute accumulations were summed up to obtain 1h accumulations considered in this paper. This dataset is used as the reference data hereinafter.

Verification software package MET

Several verification packages are used at the Hydrometcentre of Russia at present, includ-

ing VERSUS and Rfdbk packages of the COSMO Consortium. For the new high-resolution EPS verification system of the Hydrometcentre of Russia we decided to use the free MET package, which is a core of METplus verification system developed and supported to community via the Developmental Testbed Center (DTC) for use by the numerical weather prediction community (**https://dtcenter.org/community-code/metplus**). The core components of the framework include MET, the associated database and display systems called METviewer and METexpress, and a suite of Python wrappers to provide low-level automation and examples, also called use-cases. METplus is being actively developed by NCAR/Research Applications Laboratory (RAL), NOAA/Earth Systems Research Laboratories (ESRL), NOAA/Environmental Modeling Center (EMC), and is open to community contributions.

The motivation for using MET has been the availability of almost all the necessary methods in one package (Pointwise scores in the PointStat tool, grid-to-grid verification in the GridStat tool, object-based MODE method, EPS scores, etc.) and good support from MET developers. Each MET tool is set up by a configuration file and run by a bash script, thus enabling a transparent and flexible verification setup.

The MET 9.1.3 version was installed on the Roshydromet supercomputer CRAY XC40-LC under UNIX. Only MET package was used in this study, but other METplus components are being implemented now, among the most important are the METviewer for the visualization of the scores and the METplus wrappers. The primary goal of the METplus wrappers development is to provide MET users with a highly configurable and simple means to perform model verification using the MET tools. A wrapper is a Python script that encapsulates the behaviour of a corresponding MET tool.

Methods and results

Ensemble_stat tool

In the MET tool for probabilistic scores (Ensemble_stat) a large number of scores are calculated: ensemble mean, ensemble standard deviation, RPS, CRPS, Ignorance score, Probability Integral transform (PIT), Talagrand diagrams, ME and RMSE of the ensemble mean, etc. [9]. The Ensemble_stat tool also produces the spatial products: the Neighbourhood ensemble probability, NEP, and Neighbourhood maximum ensemble probability, NMEP [10, 11].

The scope of this paper is the application of the spatial approach to verification, so we are not giving examples of all standard pointwise probability scores. However, let's consider the Talagrand diagrams, also called the rank histograms [12, 13, 14], as they provide a useful means to analyzed if the EPS is enough dispersive and if it is biased. Figure 6.3.1 shows an example of the Talagrand diagram. The U-shape of the diagram in Figure 6.3.1a for lead time 2h means that the ensemble forecast is underdispersive, and the observations fall most often in the tails. In Figure 6.3.1b, for lead time 23h, most part of the ensemble members predict more precipitation than was observed, thus overforecasting bias is present. Such defaults can be partly eliminated by ensemble calibration.

Neighbourhood maximum ensemble probability, NMEP [10, 11]. In NMEP, the fraction of ensemble members is computed for which the event is occurring somewhere within the surrounding neighbourhood. The NMEP output is usually smoothed with a Gaussian kernel filter. The neighbourhood sizes and smoothing options can be customized in the configuration file. We applied the filter radius of 3 points; that is, the NMEP value at each point was averaged in the radius of three points. In Figure 6.3.1c and d, the NMEP precipitation fields are shown for two lead times. The ensemble spread is larger for longer lead time,

but it is not sufficient. Such products are useful for the forecasters, as the requirement of exact forecast at a given point is relaxed. However, they are often too smoothed out, and the local features of weather systems, in particular, of convective origin, are lost. They show the need for object-based methods [5].

MODE, Method for Object-Based Diagnostic Evaluation (MODE tool in MET) [15, 16] compares objects in gridded model and observation data. MODE was developed to mimic the subjective forecaster judgment providing at the same time the objective evaluation measure. The scores in MODE are computed as follows:

1. The objects are identified in the forecast and observation fields using the two parameters defined by a user: the threshold value for the variable under study and the convolution filter (we used the filter with a 5-points radius for this method). The objects are contiguous points where the variable exceeds the threshold.

2. The object attributes important for the user are chosen: for example, areas, centroids (geometric centres of mass of the objects), axis angle, complexity of the object contour, etc.

3. The differences in area ratio, centroid distance, angle difference, etc. are computed for each pair of objects in the fields, which are compared

4. The fuzzy logic functions F are set up to calculate the total interest, which is a measure of similarity of two objects. The fuzzy logic function controls the importance of the attribute difference. A good value of the attribute difference (for example, a small centroid distance) corresponds to the fuzzy logic function value of 1, and this fuzzy logic value decreases up to zero for the attribute difference of a useless forecast (e.g., the centroid distance is more than 200 km). Thus, all the attribute differences are transformed to values from 0 to 1 by the fuzzy logic functions. Figure 6.3.2 shows the fuzzy logic functions we choose.

5. For each attribute, a weight w is chosen and a confidence coefficient c is defined, which indicates our confidence about the attribute measuring (c=*1* was chosen here). Table 1 shows the weights w used as the basic MODE setup in this study. Weights for other attributes available in MET were set to zero.

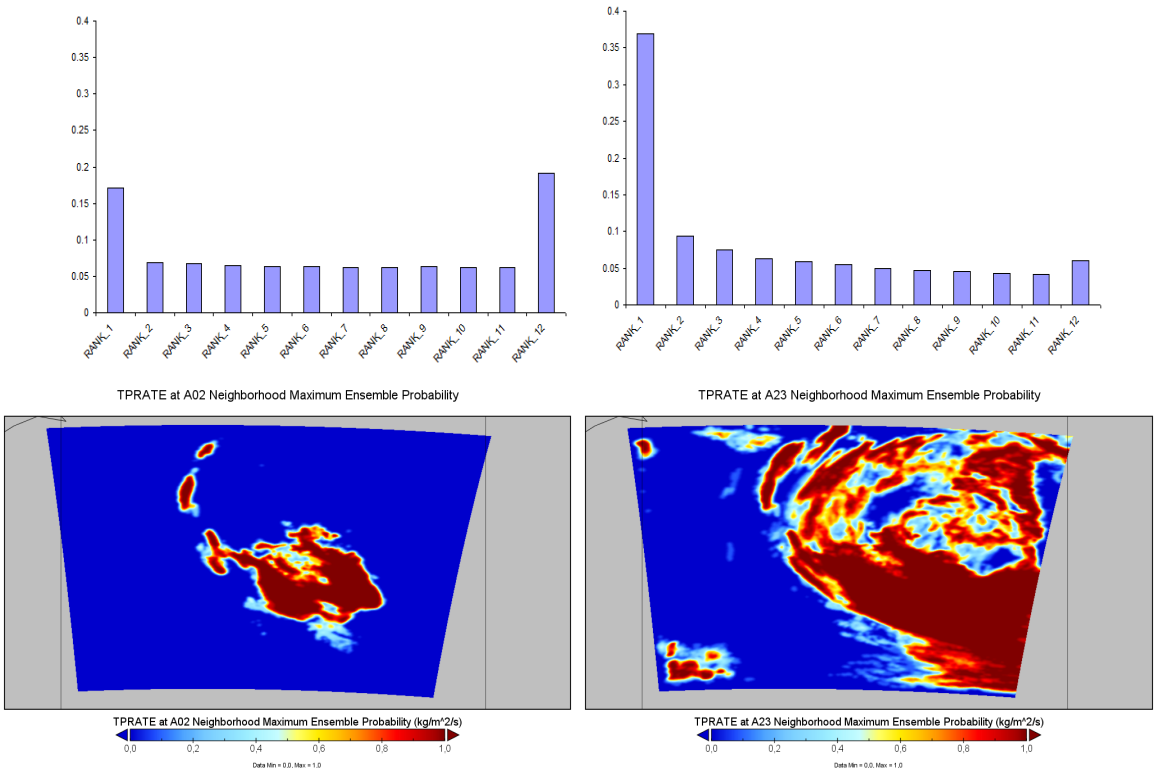


Figure 62: Products from MET Ensemble stat tool for hourly precipitation accumulations, 2021 01 July, 01-02 UTC (left column) and 2021 01 July, 22-23 UTC (right column), run from 01 July 2021, 00 UTC. Talagrand diagrams, aggregation over the forecast domain (Central Russia) (top row) and neighbourhood maximum ensemble probability, NMEP, precipitation threshold 0.1 mm/h (bottom row).

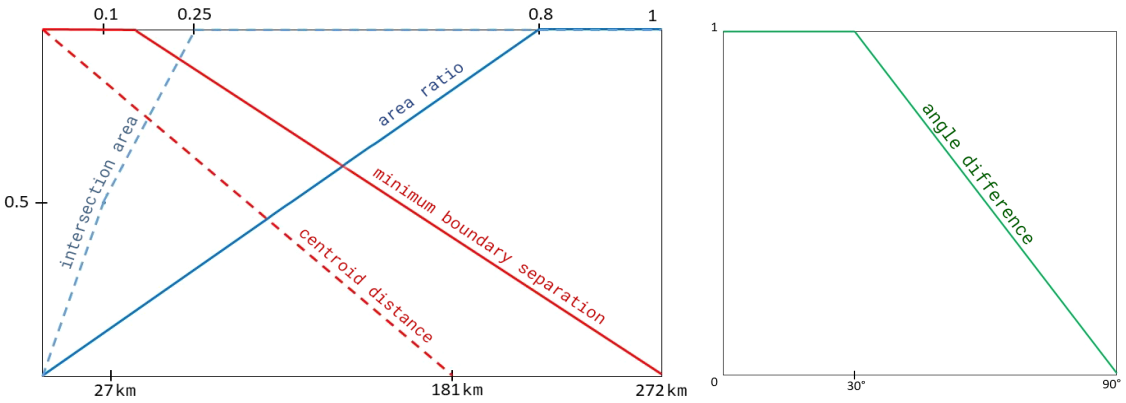


Figure 63: Fuzzy logic functions in the interval [0, 1] for the attribute differences used in this study.

6. Finally, the interest values I are calculated for each attribute i and object pair j and the total interest (TI) value is computed.

$$I_{ij} = F_{ij} w_i c_i \tag{1}$$

Fuzzy engine weights w	
Centroid distance	2.0
Minimum boundary separation	4.0
Angle difference	1.0
Area ration	1.0
Intersection area ratio	2.0

Table 16: Attributes used in the basic MODE setup and corresponding weights.

$$TI_j = \frac{\sum_i I_{ij}}{\sum_i w_i c_i} \tag{2}$$

7. The objects in MODE can be matched using a TI threshold. If the TI in a forecast-observation object pair is greater than the TI threshold, this object is considered correctly forecasted. We used the TI threshold value of 0.7 proposed by the MODE authors [15, 16]. MET has also several algorithms for merging objects within the same field, that is, object clustering. In this study, if two objects in one field happen to match the same object in the other field, then those two objects are merged.

8. Output of statistical characteristics of objects, object pairs, and clusters of objects, calculating of the median of maximum interest, MMI, the summary characteristic of forecast quality in MODE. The matrix with TI values for all pairs of objects in the forecast and observation field is composed, then the maximums in the columns and in the rows of this matrix are taken. The MMI is the median of these maximums.

The advantage of MODE is that it is a very flexible tool due to the choice of fuzzy logic functions, weights and thresholds depending on the variable and the goal of the verification. Forecast and observed objects can be matched in MODE, but it doesn't necessarily require matching. A single score, MMI, is computed, but it provides more detailed information about the object properties if needed, including the information about object misses and false alarms.

Figure 65 shows the model and reference precipitation fields (a, b) and the objects in the field of control ensemble forecast and radar composite for two thresholds, 0,1 mm/h (c-f) and 1 mm/h (g-j), for the experiment with all the attributes from (Table 16) and only the centroid distance chosen for calculating the TI. In 65 (c-j), the colours indicate matched pairs of objects, black contour indicates the merged objects within each field (i.e., object clusters), and blue colour indicates objects left unmatched (i.e., misses and false alarms). It is seen from the figure that using only the centroid distance (e, f, i, j) creates more merging. It can be more appropriate for lower precipitation thresholds producing vaster objects. However, for more intense precipitation (higher thresholds), where localization is more important, the choice of several attributes as in Table 16 (c, d, g, h) seems better.

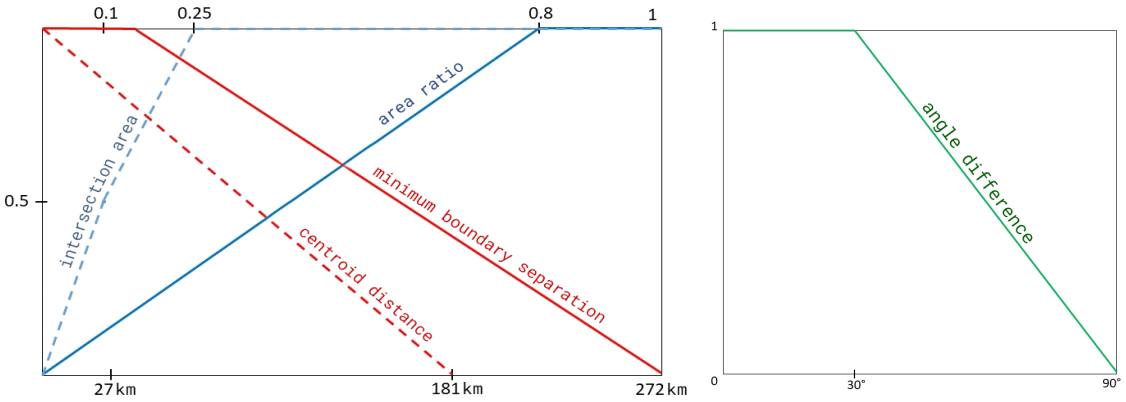


Figure 64: Fuzzy logic functions in the interval [0, 1] for the attribute differences used in this study.

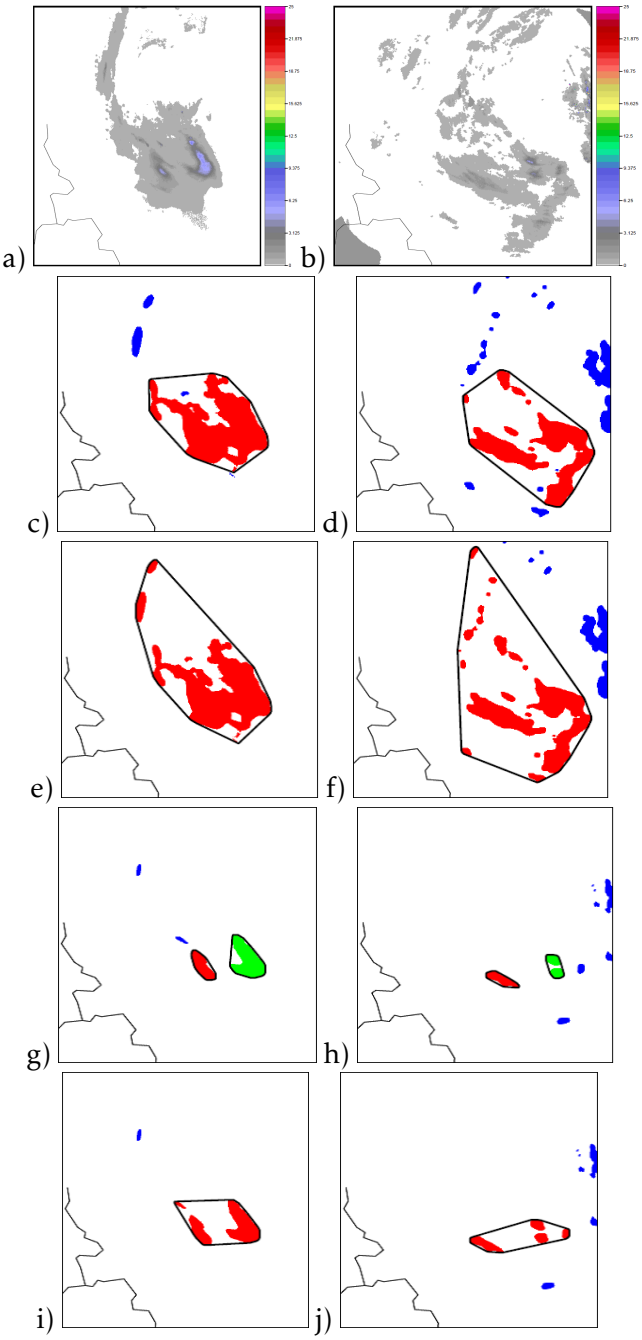


Figure 65: MODE results for hourly precipitation accumulations, 2021 01 July, 01-02 UTC, run from 01 July 2021, 00 UTC6.: (a, b) precipitation fields; (c-j) objects, (a, c, e, g, i) control ensemble member; (b, d, f, h , j) radar composite, (c-f) precipitation threshold > 0.1 mm/h; (g-j) precipitation threshold > 1 mm/h; (c, d, g, h) all attributes from Table 1 are used to calculate TI, (e, f, i, j) only centroid distance attribute is used to calculated TI.

MODE for EPS

At present, MODE is being adapted for the EPS at the Hydrometcentre of Russia. It's a development of work [7] following the approach proposed in [17, 5]. The procedure proposed in [5] consists of taking the objects from all ensemble members and constructing a hypothetical so-called ensemble "pseudomember" using the objects that are locally most representative of the ensemble distribution. These objects are obtained in several steps: 1)

A list of all objects in all forecast ensemble members is prepared; the total interest (TI) is calculated for all the pairs of the objects in this list. 2) The probabilities are calculated from the percentage of ensemble members with a matching object (matching is determined using the TI threshold as in the paragraph above). All objects are sorted by probability. Within the group of matched objects with the highest probability, the most representative object is chosen according to the highest average TI. 3) This object is added to the object list of the pseudomember. 4) The added object as well as all matching objects in its group that contributed to the probability of the added object are removed from consideration; thus a new, smaller list of objects is left. 5) Steps 2-4 are repeated until no objects remain in the list of ensemble forecast objects. Let's note that in [5] the TI calculation was modified compared to [15, 16] to multiply I components from Eq. 1, while we use the standard approach (Eq. 2) with the attributes listed in Table 1 and the matching criterion TI > 0.7 [15, 16].

Figure 6.3.4 shows the test cases with ensemble pseudomembers. It should be noted that the program is currently operating in test mode, it is planned to improve visualization, the probabilities of objects will be ordered in the legend. For case 1 (Figure 6.3.4a, lead time 2 hours), the largest object of the ensemble (object 3) is determined, which was also visible in the control member (Figure 6.3.3c). The probability of object 3 is 100%. This means that it is found in each of the ensemble members. Objects 1, 2, 5 also have a 100% probability. Part of the selected objects is not visible in the chart, this is due to the fact that the selected objects may be small or overlap with other objects. A possible way to solve this problem is to increase the weight of attributes related to object position matching. Another possible way is to combine close objects into clusters before running the pseudomember calculation (MET MODE provides several clustering algorithms).

Figure 6.3.4b shows the pseudomember for case 2 (lead time 23 hours). It consists of a very large number of small objects. However, the largest object of the control member was not included in the pseudo-member, because it is not the most representative of the contribution of the area attribute. **This demonstrates the added value of the ensemble, since the pseudomember fits the observational data better than the control term** containing unrealistically large precipitation area. The selected objects have different probabilities, which indicates a spread over the ensemble.

In Figure 6.3.4c, only five objects are selected, and only one of them has a 100% probability of appearing in the ensemble. Other pseudomember objects are close to each other and practically form a single object. The most common of these is object 2, which has a 63.64% probability of appearing in an ensemble member, meaning that it occurs in more than half of the ensemble members. The ensemble underestimates precipitation.

Figure 6.3.4d shows the pseudomember for case 4 (lead time 23 hours). Many small objects characterize it, some of them overlapping. In total, 74 objects were identified for case 4. Note that according to observations, there are many separate small objects, so the pseudo term is closer to reality than the control one. Another conclusion from this case is that it may be useful at the initial stage of the analysis to filter out objects that are less than a certain area threshold in the forecast and observation fields.

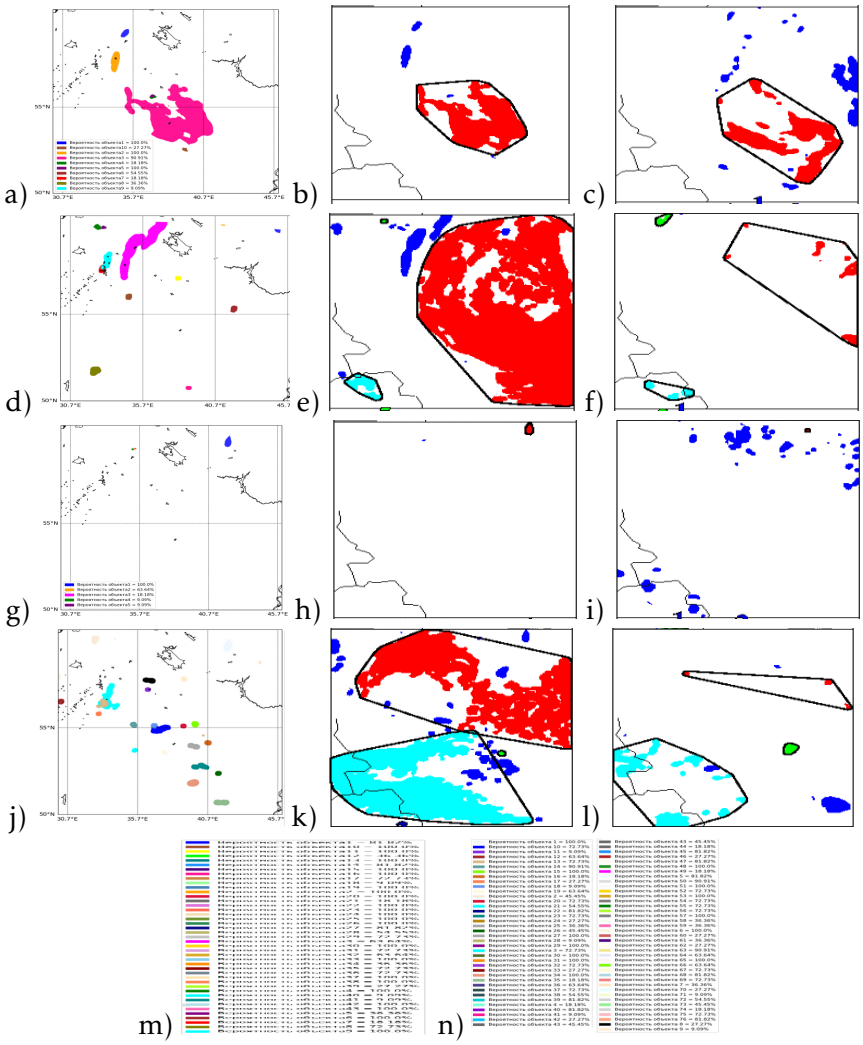


Figure 66: Ensemble object pseudomember (left panel: a, d, g, j) with probabilities of the objects, ensemble control forecast objects (centre panel: b, e, h, k) and observation objects (right panel: c, f, i, l); precipitation threshold > 0.1 mm/h. (first row: a-c) 2021 01 July, 01-02 UTC, run from 01 July 2021, 00 UTC; (second row: d-f) 2021 01 July, 22-23 UTC, run from 01 July 2021, 00 UTC; (third row: g-i) 2021 02 July, 01-02 UTC, run from 02 July 2021, 00 UTC; (fourth row: j-l) 2021 02 July, 22-23 UTC, run from 02 July 2021, 00 UTC.

Note: Figure 66m is the legend with object probabilities for Fig. 66d and Fig. 66n is the legend for Fig. 66j, given outside the plots because of a large number of the objects in these test cases.

The pseudomember will be used later to obtain MODE estimates for the ensemble forecast using a set of threshold values for the probability of objects. It is made in the following way: For a fixed probability threshold, only objects with probabilities above this threshold are selected, and the rest are discarded; then MODE is run in the usual deterministic mode. Thus, a set of MODE scores is formed for each probability threshold.

It was found in this analysis that in some cases overlapping of the selected objects occurs. In our opinion, this is an undesirable feature. To avoid it, more experiments with attribute weights are underway.

Conclusions and plans

The high-resolution ensemble verification system is being developed at the Hydrometcentre of Russia based on the free MET package. It includes both standard and spatial verification techniques. Verification is performed on the precipitation test cases up to now. The MET package is found to be a highly configurable and flexible tool. At present, accumulation and aggregation of the scores for larger number of forecast cases and for the variables besides precipitation are under preparation. The experiments will be continued with MODE method applied to deterministic and ensemble forecasts to find the optimal combination of tuning parameters for different weather situations.

The ensemble forecasts examined in this paper are underdispersive as they were performed within the research of purely model uncertainty and therefore no perturbations of initial and boundary conditions were introduced. Further, we plan to apply the verification system to ensemble forecasts with perturbations of initial and boundary conditions downscaled from the global ensemble prediction system ICON-EPS. We will also extend the verification period.

References

[1] Leutbecher M and Palmer T N 2008 Ensemble forecasting *J. Comput. Phys.* 227 Issue 7 3515–39 https://doi.org/10.1016/j.jcp.2007.02.014

[2] Palmer T 2019 The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Q. J. R. Meteorol Soc.* **145** (Suppl. 1) 12–24 https://doi.org/10.1002/qj.3383

[3] Montani A, Cesari D, Marsigli C and Paccagnella T 2011 Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges, *Tellus A: Dyn. Meteorol. Oceanogr.* **63:3** 605–624 https://doi.org/10.1111/j.1600-0870.2010.00499.x

[4] Gilleland E, Ahijevych D A, Brown B G and Ebert E E 2010 Verifying Forecasts Spatially *Bull. Am. Meteorol. Soc.* **91** 1365–73 http://dx.doi.org/10.1175/2010BAMS2819.1

[5] Johnson A, Wang X, Wang Y, Reinhart A, Clark A J, and Jirak I L 2020 Neighborhood- and Object-Based Probabilistic Verification of the OU MAP Ensemble Forecasts during 2017 and 2018 Hazardous Weather Testbeds *Weather Forecast.* **35**, 169–191, https://doi.org/10.1175/WAF-D-19-0060.1

[6] Kiktev D, Joe P, Isaac G A, Montani A, Frogner I, Nurmi P, Bica B, Milbrandt J, Tsyrulnikov M, Astakhova E, Bundel A, Bélair S, Pyle M, Muravyev A, Rivin G, Rozinkina I, Paccagnella T, Wang Y, Reid J, Nipen T and Ahn K 2017 FROST-2014: The Sochi Winter Olympics International Project *Bull. Am. Meteorol. Soc.* **98** 1908–29 https://doi.org/10.1175/BAMS-D-15-00307.1

[7] Bundel A, Muraviev A, The contiguous rain area (CRA) method application for the Caucasus and Alpine regions 2017 *Research activities in atmospheric and oceanic modelling. CAS/JSC Working Group on Numerical Experimentation. Report No. 47. WCRP Report No.12/2017 ed E. Astakhova (WMO, Geneva) 10-03–04*

[8] Zängl G Reinert D Ripodas P and Baldauf M 2015 The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Q. J. R. Meteorol. Soc.* **141** 563–579 https://doi.org/10.1002/qj.2378

[9] Wilks D S *Statistical Methods in the Atmospheric Sciences* 2019 (Elsevier) pp 816

[10] Schwartz C S, Kain J S, Weiss S J, Xue M, Bright D R, Kong F, Thomas K W, Levit J J, Coniglio M C and Wandishin M S 2010 Toward Improved Convection-Allowing Ensembles:

Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Ensemble Membership, *Weather Forecast.* **25** 263–280 https://doi.org/10.1175/2009WAF2222267.1

[11] Schwartz C S and Sobash R A 2017 Generating Probabilistic Forecasts from Convection-Allowing Ensembles Using Neighborhood Approaches: A Review and Recommendations, *Mon. Weather Rev.* 3397–3418 https://doi.org/10.1175/MWR-D-16-0400.1

[12] Talagrand O 1997 Statistical consistency of ensemble prediction systems *Discussion paper, SAC Working Group on EPS*, ECMWF

[13] Hamill T M and Colucci S J 1997 Verification of Eta-RSM short-range ensemble forecasts *Mon. Weather Rev.* **125** 1312-27
https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2

[14] Hamill T M and Colucci S J 1998 Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts *Mon. Weather Rev.* **126** 711-724
https://doi.org/10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2

[15] Davis C, Brown B and Bullock R 2006 Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas, *Mon. Weather Rev.* **134** 1772–1784

[16] Davis C A, Brown B G, Bullock R and Halley-Gotway J 2009 The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program *Weather Forecast.* **24** Issue 5 1252-67 https://doi.org/10.1175/2009WAF2222241.1

[17] Johnson A and Wang X 2012 Verification and Calibration of Neighborhood and Object-Based Probabilistic Precipitation Forecasts from a Multimodel Convection-Allowing Ensemble *Mon. Weather Rev.* **140** 3054-77 https://doi.org/10.1175/MWR-D-11-00356.1

6.4 DIST methodology tuned on high-threshold events for flash floods forecast evaluation

Maria Stefania Tesini, ARPAE-SIMC

Introduction

This task proposed to explore and highlight the suitability of an evolution of the DIST methodology *(see Marsigli, C., Montani, A., and Paccagnella, T.: A spatial verification method applied to the evaluation of high-resolution ensemble forecasts, Meteorol. Appl., 15, 125–143, 2008)* for the verification of HIW, such as high precipitation over catchment areas used operationally for issuing Civil Protection alerts

The proposed methodology has been developed as a spatial method for the verification of heavy precipitation issued at high resolution. In fact, it permits the use of a high-resolution rain-gauges network, but gridded observations, such as radar precipitation analysis, can be used as well. The main advantage of this approach is that no precipitation analysis is required and information about localized maxima of precipitation can be considered, as well as the variability of the precipitation field inside the area of interest.

Similarly, all the grid points that belongs to the selected area are considered, in this way the ability of the model in reproducing high precipitation events, even if with some possible positioning errors, is evaluated.

Verification results can be used directly to interpret how to use the forecast system and to decide in which situations one system is better than another.

The verification system

It is an evolution of DIST, a spatial verification method based on the verification of the precipitation distributions within boxes of selected size (Neighbourhood obs – Neighbourhood fcs). In DIST methodology, the verification domain is subdivided into boxes, each of them containing a certain number of observed and forecast values.

For each box, several parameters of the distribution of both the observed and forecast values falling in it can be computed (mean, median, percentiles, maximum).

Verification is then performed using a categorical approach, by comparing for each box one or more parameters of the forecast distribution against the corresponding parameters of the observed distribution, using a set of indices.

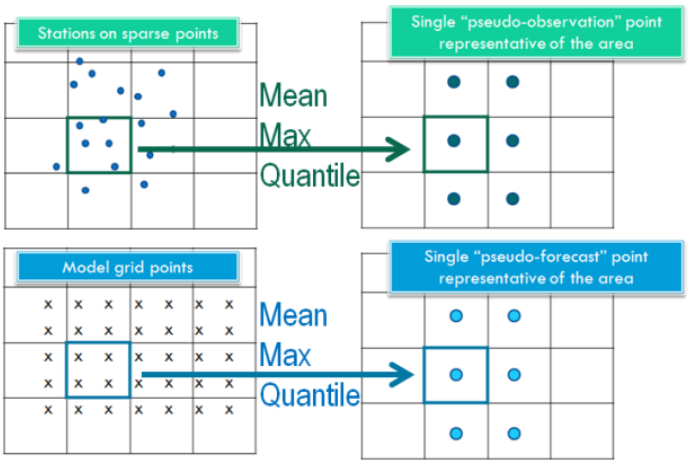


Figure 67: Ensemble Original DIST methodology.

In the evolution of the methodology, squared regular boxes are replaced with catchment areas. One of the main reason of this choice is the need to reduce problems related to complex terrain, e.g. if a ridge of a mountain divides the box this can give misleading results combining upwind and downwind situation.

A second aspect, no less important, is the possibility of communicating the results more easily and directly to end users (e.g. meteorologists or hydrology) because the scores can be provided on each catchment area.

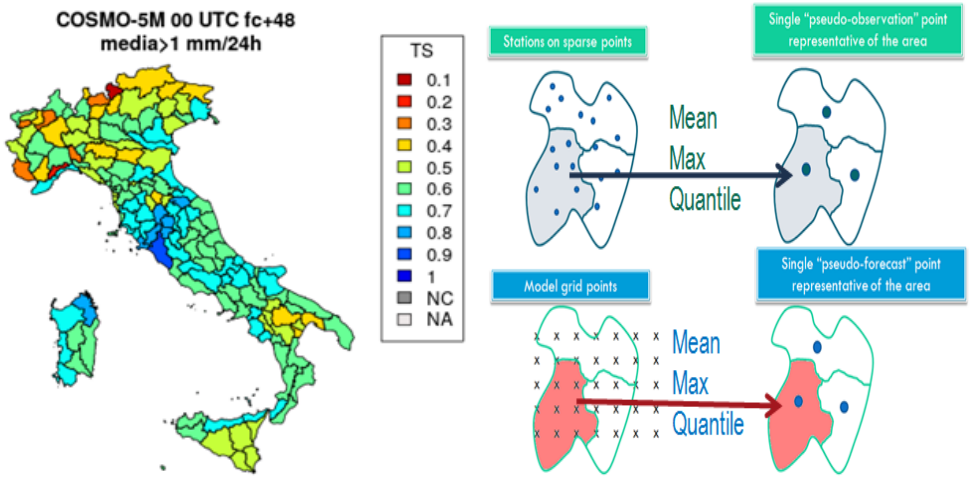


Figure 68: Example of score (TS) on the Italian catchment areas (left) and Evolution of DIST methodology (right).

The new methodology has been validated over Italy comparing results from DIST original "squared boxes" and from new catchment areas considering the maximum value exceeding some thresholds in each box or area.

The improvement in the scores obtained using the catchment area as a reference for verification seem to support the choice made, in particular by reducing the number of false alarms and increasing the Success Ratio for all the considered models and thresholds.

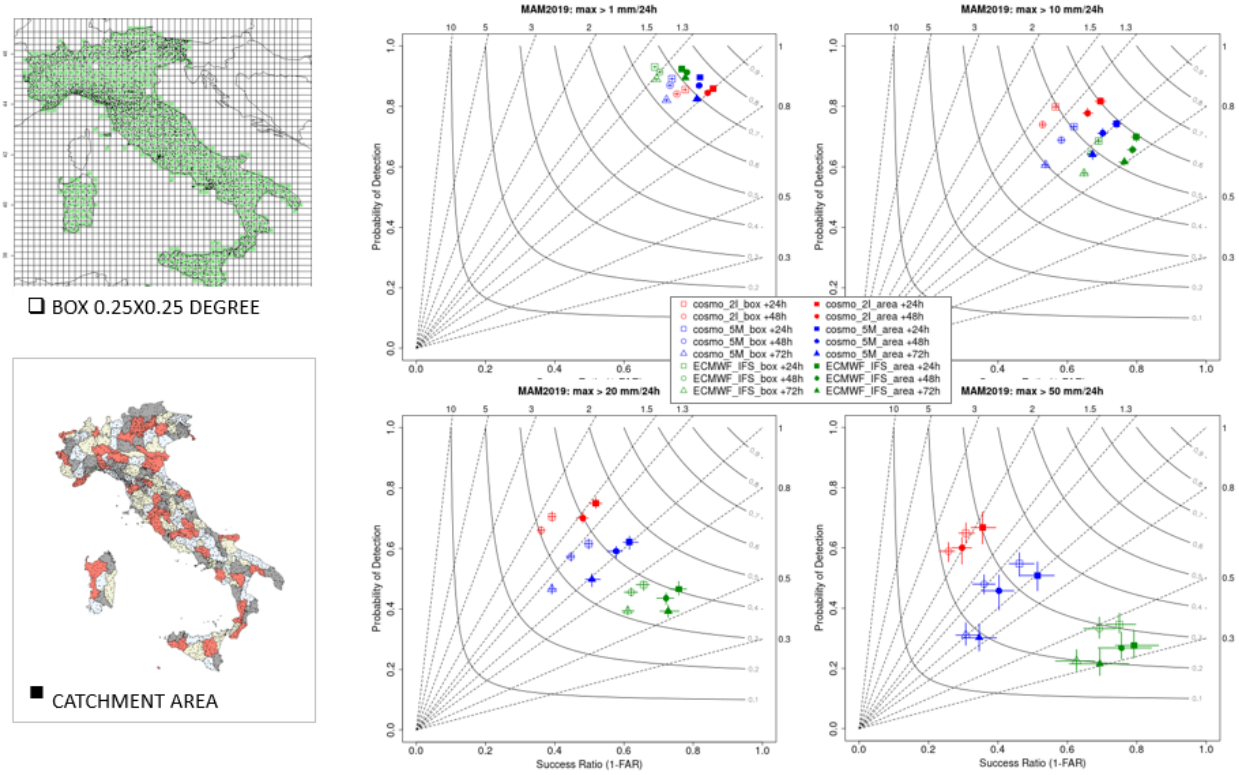


Figure 69: Results comparing maximum value exceeding thresholds of 1 mm/24h (top left), 10 mm/24h (top right), 20 mm/24h (bottom left) and 50 mm/24h (bottom right) during MAM2019 over Italy. The colour of the symbols represents the different models (red for Cosmo-2I, blue for Cosmo-5M and green for IFS-ECMWF. Filled symbols are for scores evaluated on the catchment areas, empty ones for those on squared boxes.

Operational use of DIST: interpretation of the results

One of the main goals of this verification methodology is to provide to end users results that can be used directly to interpret how to use the forecast system and to decide in which situations one system is better than another.

Considering different parameter of the precipitation distribution in each area it is possible to focus the attention on some characteristics of the precipitation field:

1. Average: it can be used to investigate the ability of models in reproducing different amounts of precipitation over each area. Hydrologist are very interested in this information.
2. Maximum: the use of the maximum of precipitation over the areas can provide some information on high precipitation, even if not in the correct location but in the neighbourhood, represented by the catchment area.
3. Median & Maximum: the combination of a condition on the median and one on the maximum of precipitation can separate high localized precipitation from extensive precipitation.

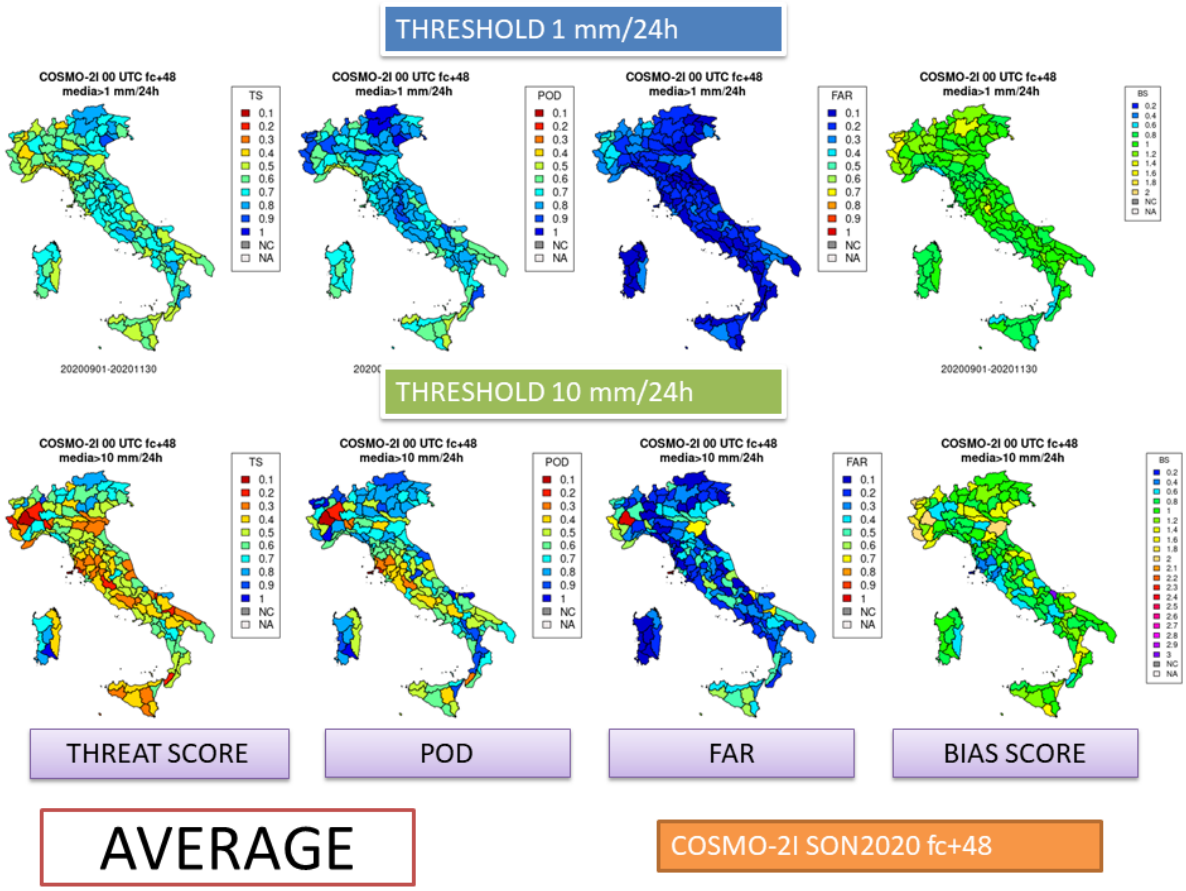


Figure 70: Example of maps of different scores for the event defined as average precipitation in the area greater than 1 mm/24h (in the upper panel) or 10 mm/24h (in the lower panel). Differences can be noted both between the various areas and on the same area for different thresholds.

The comparison between the results of the verification using the average or maximum value over the area allows to highlight the different behaviour of the models: in many cases lower resolution models have better performance considering the mean values, but they tend to underpredict precipitation maxima. On the other side higher resolution models such as convection permitting models are less performant on predicting average values, but they are able to forecast higher values of precipitation, at the expense of a large number of false alarms.

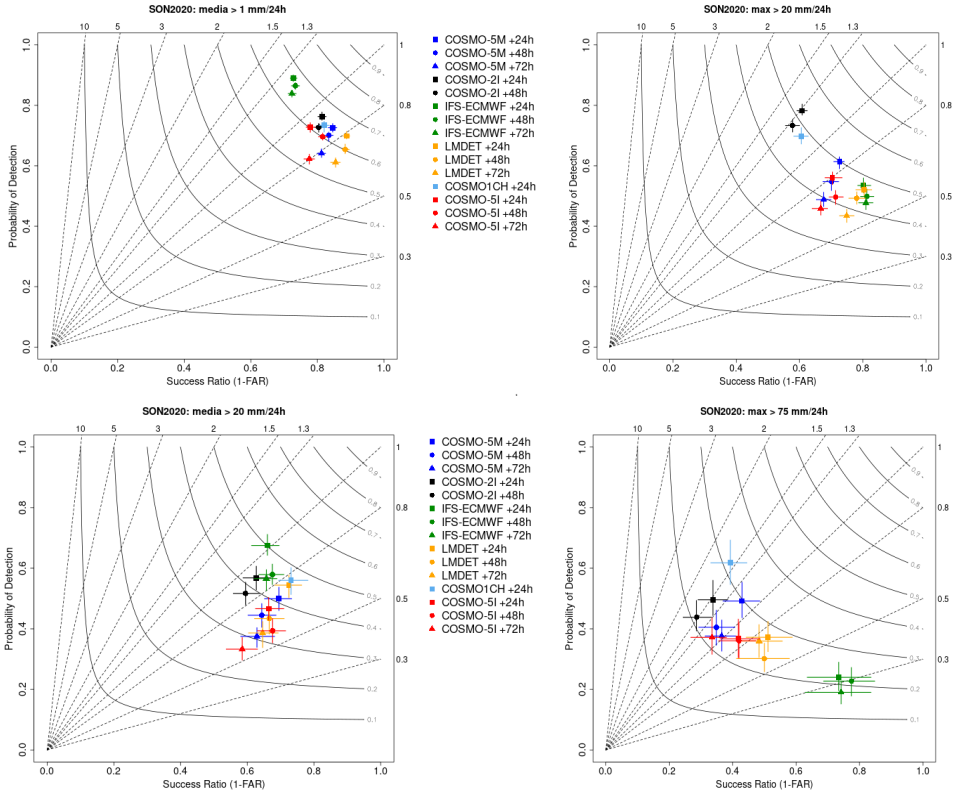


Figure 71: Performance Diagrams showing results of verification for SON2020 over Italian catchment areas for different models and forecast steps. In the left panels the mean value exceeding the threshold of 1 mm/24h (top) and 20 mm/24h (bottom) are considered, while on the right panels are reported the results concerning the maximum value exceeding the threshold of 20 mm/24h (top) and 75 mm/24h (bottom).

The user can then form his own opinion on the performance of the model based on the type of use he has to make of it. For example, if the interest is aimed at issuing alerts for the possibility of high precipitation, the choice to give more credit to the model that has the best results on the average precipitation could lead to numerous missed alarms. In this case it would be preferable to take the higher resolution model into consideration.

Furthermore, we investigated the possibility of characterizing the distribution of precipitation in the area.

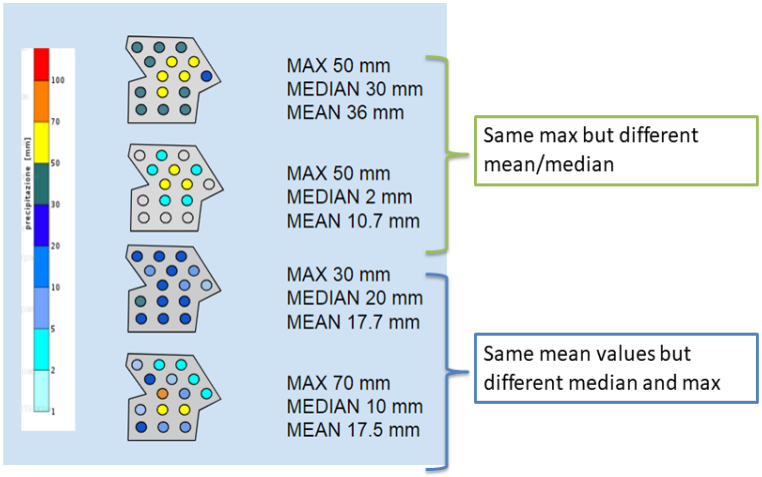


Figure 72: Schematized example of precipitation distribution over an area. The dots represent the rain-gauges or the grid-points of the model, the colours the amount of rain.

We are not interested in the exact position of the maxima, the DIST method is in fact used to minimize spatial errors, but the idea is to discriminate, as a first approximation, between high but localized precipitation and widespread rainfall. For this purpose, we tested the use of the combination of two conditions: one on the maximum of precipitation and one on the median.

For example, a precipitation with a maximum on the area greater than 50 mm/24h can be due to different scenarios, which is possible to represent imposing a condition on the median:

1. intense and widespread precipitation: in at least half of the points in the area it rained more than 30 mm/24h (median > 30 mm/24h) with at least one (the maximum) greater than 50 mm /24h.
2. intense precipitation but not extended to the whole area: in half the points of the area it rained less than 20 mm/24h (median <20 mm/24h), while in the other half of the points of the area it rained more than 20 mm/24h with at least one point (the maximum) greater than 50 mm/24h.
3. intense but more localized precipitation: it rained in half of the points in the area less than 10 mm/24h (median <10 mm/24h),), while in the other half of the points of the area it rained more than 10 mm/24h with at least one point (the maximum) greater than 50mm/24h.

To support the use of the median/maximum combination to distinguish the various precipitation scenarios, since we know that most of the high localized precipitation are due to convective events, we considered a dataset of observed precipitation over the eight alert areas of the Emilia-Romagna region and the corresponding lightning data. Different conditions on median an maximum had been imposed to separate the scenarios, then, for each different scenario, the events were classified on the base of the presence of lightning or not, with the assumption that if there is lightning the precipitation is of convective type.

The dataset was composed of 270 day of observed precipitations for the period March-November 2015, described by median and maximum on the 8 alert areas of Emilia-Romagna

region, and the corresponding total number of lightning per day over each area, simplified in lightning/no-lightning.

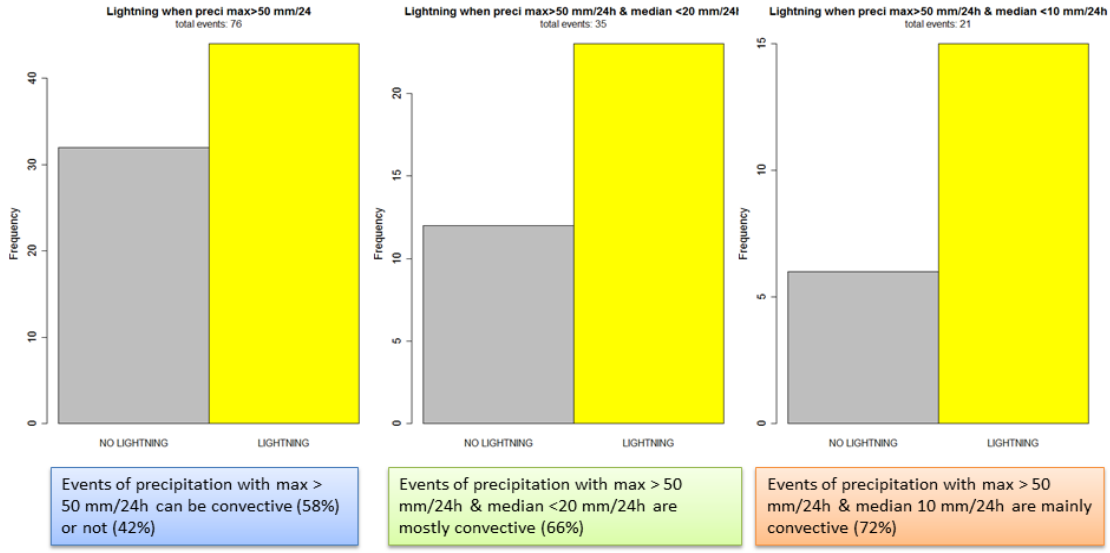


Figure 73: Classification of observed precipitation events on the base of conditions on maximum and median of precipitation over the eight alert areas of Emilia-Romagna region, for a period of 270 days (March-November 2015). Each category has been further stratified based on the presence of lightning.

Out of a total of 2160 cases considered, 1571 were rain events but in only 76 the maximum of precipitation exceeds the 50 mm/24 thresholds. Considering these high precipitation events with no condition on the median, the 58% were convective and the 42% not-convective (i.e. with lightning or not) and it's difficult to distinguish between cases of convection or not. But if the condition "median less than 20 mm/24h" is imposed, the events associated with the presence of lightning are the 66% of the total. The number of convective events then becomes 72% when the condition "median less than 10 mm 24/h".

These results confirm that using joint use of conditions on maximum and median can be a good approximation to select high localized precipitation that are mainly due to convection.

By applying this type of classification in the verification activity, it is possible to evaluate the behavior of the models in the reproduction of different precipitation scenarios, highlighting, in broad terms, the meteorological situation in which the models perform better or worse.

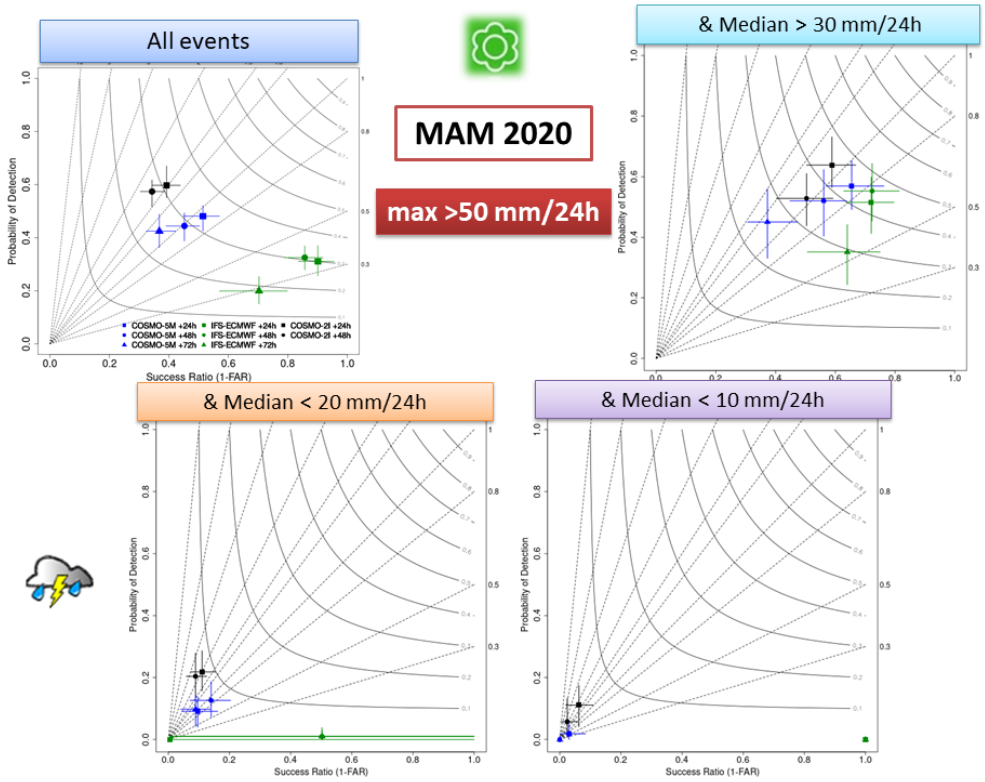


Figure 74: Performance Diagrams showing results of verification for MAM2020 over Italian catchment areas for different models and forecast step considering the events with maximum precipitation exceeding 50 mm/24h. In the top left panel all the events are considered: in the top right panel only those with median greater than 30 mm/24h, representing a scenario of intense and widespread precipitation, in the bottom panels those with median less than 20 mm/24h (left) and 10 mm/24h (right), representing scenarios of high of localized precipitation (e.g. mainly thunderstorms).

For example, referring to the results of MAM2020 in which the maximum precipitation in the area exceeds the threshold of 50 mm/24h, it is possible to attribute the high number of false alarms of the COSMO models or of the misses of IFS-ECMWF to a bad representation of localized convection, while considering intense but diffuse precipitation the scores are in general better.

This type of information can provide the user with a more complete picture of the forecasting system.

Communication of the results

Reports of verification results obtained applying the DIST methodology illustrated in this work are produced on a seasonal basis internally for Arpae and Civil Protection usage. Several COSMO models with different resolution and IFS-ECMWF are considered over the Italian catchment areas using the dataset of rain-gauges provided from the National Department of Civil Protection.

6.5 LPI verification and correlation of convective events with microphysical and thermodynamical indices

F. Gofa, D. Boucouvala, HNMS

Approach

The distribution analysis of several convective events in both space and time will allow lightning/thunderstorm regimes to be determined. Lightning Potential Index (LPI) is a measure of the potential for charge generation and separation that leads to lightning flashes in convective thunderstorms and can be calculated from COSMO model. While the connection between cloud microphysics and lightning seems apparent, the common indices used for forecasting thunderstorms and the potential for lightning usually rely on stability and thermodynamical indices (e.g., CAPE).

An effort will be given to correlate LPI with observed lightning. In this way, it will be evaluated if for Greek territory LPI can be useful parameter for predicting lightning as well as a tool for improving weather forecasting of convective storms and heavy rainfall. Statistical evaluation of LPI forecasts with traditional dichotomic scores as well as with SAL spatial method on selected intense convective events will be also performed by comparing gridded lightning data with model forecasts. LPI will be evaluated (optimum upscale window) over Greece on certain events, as a useful parameter for predicting lightning as well as a tool for improving weather forecasting of convective storms and heavy rainfall.

Lightning formation

The microphysical processes that lead to the formation of precipitation particles are involved in charge separation and the build-up of electric fields in convective clouds.

The non inductive mechanism, involves rebounding collisions between graupel particles and cloud ice crystals and requires the presence of supercooled liquid water. Lightning Potential Index (LPI) is a measure of the potential charge separation that leads to lightning flashes in convective thunderstorms (Yair et al. 2010, JGR). It is calculated from model simulated updraft and microphysical fields within the charge separation region of clouds between (0^o C and - 20^o C), where the non-inductive mechanism involving collisions of ice and graupel particles in the presence of supercooled water is most effective (Saunders, 2008).

LPI is defined as the volume integral of the total mass flux of ice and liquid water within the "charging zone" in a developing thundercloud. The LPI (J kg-1) and is defined as:

$$LPI = \frac{1}{V}\iiint \in \omega^2 dxdydz$$

$$\in = 2(Q_i Q_l)^{0.5} / (Q_i + Q_l)$$

Where V is the volume of air in the layer between 0^oC and -20^oC, w is the vertical wind component (ms-1) and q_s, q_i and q_g are the model-computed mass mixing ratios for snow, cloud ice, and graupel respectively (in kg-1). ϵ is a dimensionless number that has a value between 0 and 1 and is defined by the formula above.

Where: Q_l is the total liquid water mass mixing ratio and Q_i is the ice fractional mixing ratio (kg/kg) defined by,

$$Q_i = q_g \left[\left((q_s q_g)^{0.5} / (q_s + q_g) \right) + \left((q_i q_g)^{0.5} / (q_i + q_g) \right) \right]$$

ε is a scaling factor for the cloud updraft and attains a maximal value when the mixing ratios of supercooled liquid water and of the combined ice species (the total of cloud ice, graupel, and snow) are equal.

Calculation of the LPI from the cloud-resolving atmospheric model output fields can provide maps of the microphysics-based potential for electrical activity and lightning flashes.

Methodology

Model setup: LPI can only be calculated if you run model with the graupel microphysics (itype_gscp=4) or the 2-moment microphysics. Results for LPI are only meaningful in convection resolving mode, i.e., deep convection parametrization switched off and grid spacing smaller or equal to 4 km. LPI is a column integral involving the square of the vertical velocity and the presence of graupel (=rimming process) and other ice hydrometeors at the same locations. It needs explicitly simulated convective cells with realistic updraft speeds.

The COSMO-GR4 LPI forecasts were used with 0.04 deg resolution forecasts (not a operational product) as well as CAPECON outputs, while other indices were calculated from model outputs. This serves as the original resolution of the analysis performed. Then aggregated forecast and observations gridded format with multiple of the original space resolution are calculated through scripts that were developed.

Forecasts gridded fields: For the original resolution (0.04), the LPI value of each grid point is checked, and if it is higher than the value of 0.3 (see table below) , a value of 1 is given to the specific grid point. Next, grids with increased (multiple) resolution based on the original dimensions are created (e.g., 0.04x2, 3, . . . , 20). For each new grid cell or each new grid, the MAX LPI value of the 3x3 points is assigned.

Observations: For all new grids with resolution from 0.04deg up to 20x0.04deg, a lat-lon based check is performed in the boundaries of each grid cell, for the existence of lightning observations and a value of 1 is assigned to that grid point for positives checks or else a value of zero.

Statistics:

To statistically evaluate LPI forecast performance the following are applied:

1. Direct comparison of obs-fcs gridded values and calculation of contingency table properties.
2. SAL methodology for steady LPI threshold of one (lightning existence).

Thermodynamical indices

Stability indices were calculated using temperature and relative humidity profiles from the COSMO-GR4 model forecasts. The formulas used for the estimation of the various indices in this analysis are specified below.

K index (KI)

It calculates the thunderstorm potential based on the vertical temperature lapse rate between 850 and 500 mbar pressure levels, moisture content at 850 mbar pressure and moist layer depth at 700 mbar pressure (George 1960).

$KI = (T_{850}\text{-}T_{500})+Td_{850}\text{-}(T_{700}\text{-}T_{700})$,

with the suffix values indicating the pressure level.

The critical values of KI index indicating thunderstorm activity (Johnson 1982) are given below:

KI (K)	**Thunderstorm chances**
under 288	0% Chance
In the middle of 288 and 293	20% chance
In the middle of 294 and 298	20-40% possibility for little thunderstorms
In the middle of 299 and 303	40-60% possibility for little to medium thunderstorms
In the middle of 304 and 308	60-80% possibility for heavy thunderstorms
In the middle of 309 and 313	80-90% possibility for severe thunderstorm event
above 313	Over 90% possibility for thunderstorm event

Total Totals Index (TTI)

The TTI is procured by basic deduction among temperature and dew point temperature values at 850 and 500 hpa pressure levels (Miller 1967).

Cross totals, $CT = Td_{850} - aT_{500}$; Vertical totals, $VT = T_{850} - T_{500}$

Total Totals Index, $TTI = CT + VT = T_{850} + Td_{850} - 2T_{500}$

The critical threshold values of TTI parameter (Miller 1972) are given below:

ITTI values (K)	**Thunderstorm possibility**
Ranging between 44 and 45	Possibility for small thunderstorm activity
Ranging between 46 and 47	Possibility for moderate thunderstorm activity
Ranging between 48 and 49	Possibility for moderate to severe range of thunderstorm activity
Ranging between 50 and 51	Possibility for heavy thunderstorm activity
Ranging between 52 and 55	Possibility for scattered thunderstorm activity
above 55	Possibility for severe thunderstorm activity

Improved total totals index

The improved total totals index is obtained by the average of the temperatures at surface (at 2 m), the 925hpa and the 850hpa pressure levels (Miller 1967).

$ITTI = (2mT+ Td_{925}+T_{850})/3 +(2mTd+ Td_{925}+Td_{850})/3 - 2T_{500}$ -

The threshold for thunderstorm occurrence is usually seen at 57 K.

Humidity Index (HI)

It is obtained by calculating the availability of water vapour at 850, 700 and 500 hPa pressure levels. The importance of relative humidity as the major component needed for the severe thunderstorm activities is being estimated by this index.

HI= (T_{850}-Td_{850})+(T_{700}-Td_{700})+(T_{500}-Ts_{500})

When HI values lies less than or equal to 30K, high possibility for thunderstorm occurrence has been noticed on that region.

Convective available potential energy (CAPE)

The buoyant energy required to accelerate an air parcel vertically is referred to as CAPE. The sum of positive buoyant energy from the level of free convection to the equilibrium level can be used to measure it (Moncrieff and Miller 1976).

$$CAPE = \int_{x}^{y} g[\frac{TV_{parcel} - TV_{env}}{TV_{env}}]\, dz$$

Where TV_{parcel} represents the parcel's virtual temperature and TV_{env} represents the virtual temperature of environment respectively. x and y denote the level of free convection and neutral buoyancy. The critical values of cape parameter (Grieser 2012) are:

CAPE (IN J/KG)	Thunderstorm chances
UNDER 300	no energy for convection
FROM 300 TO 1000	Poor potential for weak convection
FROM1000 TO 2500	moderate potential for convection
GREATER THAN 2500	strong potential for convection

Selection of intense precipitation events

For the application of the methodology, eight test cases with significant convective precipitation amounts around Greece were analyzed, thus only three of them were proved to be significant and presented with respect to the LPI values forecasted.

1. Test Case 1: 15 Nov 2017
2. Test Case 2: 12 Nov 2019
3. Test Case 3: 24 Nov 2019

Other cases analyzed were: 10/07/2019, 07/12/2020, 08/08/2020, 02/06/2018, 03/10/2019. Below the synoptic description for weather situation is provided.

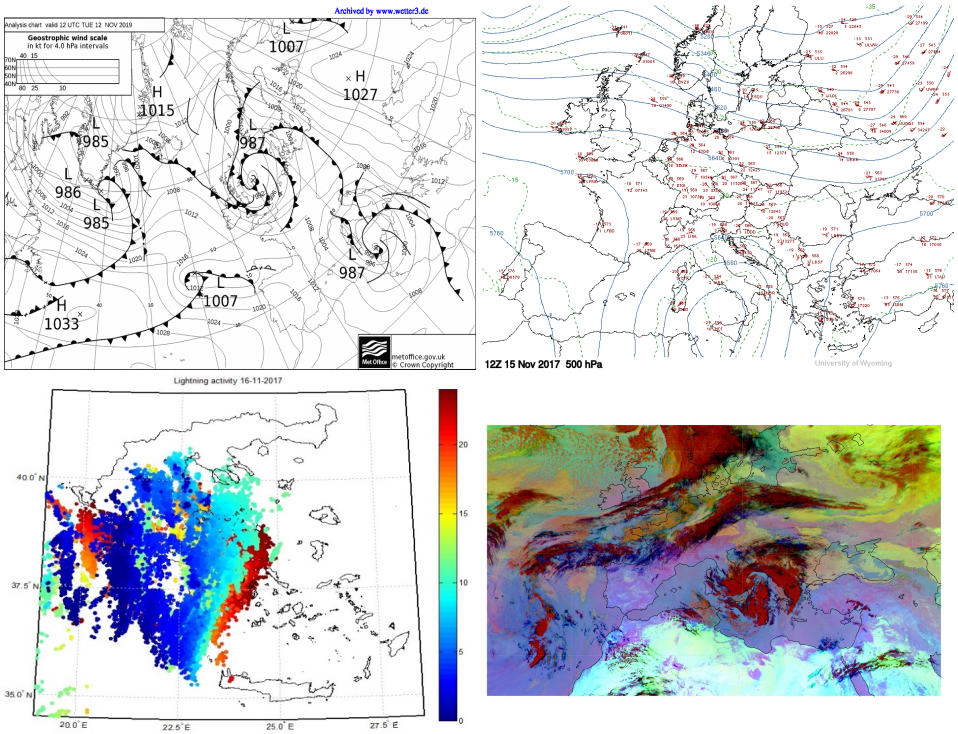


Figure 75: Upper row: MSLP maps during the event. Lower row: accumulated observed lightnings (left) and RBG air masses.

1. **Test Case 1**

A cut-off low in Tunisia over upper troposphere on 11/11/2017 associated with a low over Syrti Gulf which caused severe thunderstorms over Central Mediterranean, moved northeastwards and on 13/11 influenced initially western Greece and gradually east parts, mainly Attica, Cyclades, Crete and Dodecanese with heavy phenomena. In addition, on 13/11 a second deep low over Genoa Gulf (995hpa) transferred polar air masses over Southern Italy. On 15/11 the low expanded and moved over Central Mediterranean. Over the warm sea of Ionian, the cold air destabilized. Due to weak wind shear, a cyclone (Medicane) was formed. Heavy rainfall and flooding caused severe damage over Western Attica (Fig. 75).

1. **Test Case 2**

Deep barometric low with frontal activity over South Italy moved north eastwards leading to strong gale southerly winds (9 Beaufort), heavy rain, thunderstorms and electrical discharge all over Greece (except Dodecanese). Flooding over Attica and Crete were reported while Ionian islands suffered from severe damages especially Corfu and Cefalonia (Fig. 76.)

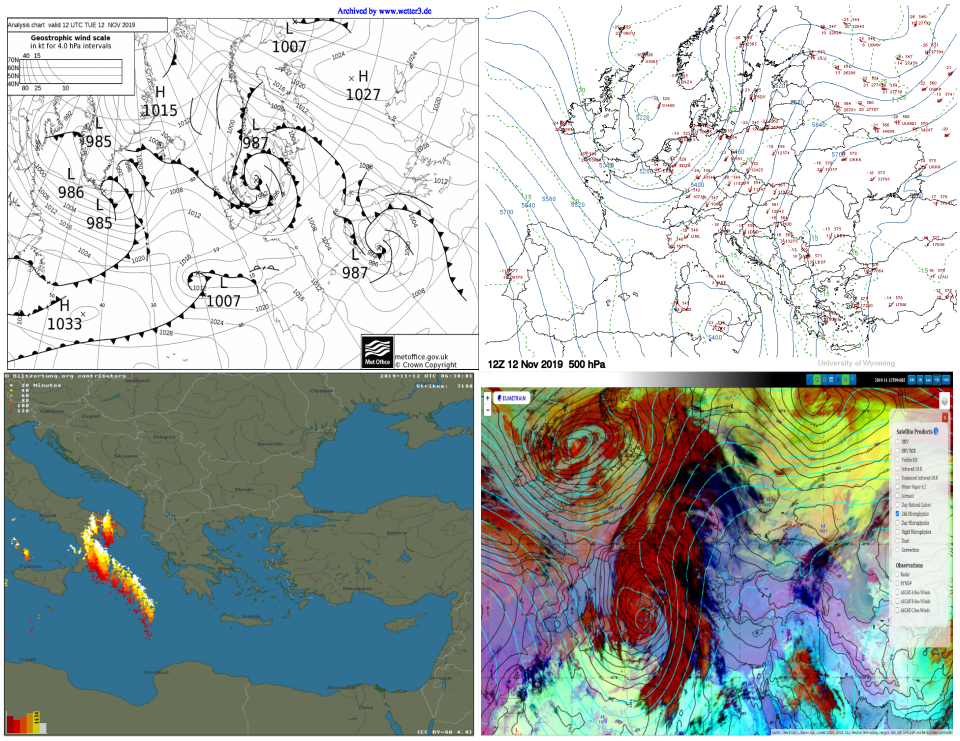


Figure 76: Upper row: MSLP maps during the event. Lower row: accumulated observed lightnings (left) and RBG air masses.

1. **Test Case 3**

Deep low over Italy moved eastwards and produced a cold front over Ionian Sea which influenced all the country of Greece with severe damages due to heavy rainfall. Flooding were reported in South Attica, Rhodes, Central Macedonia and East Aegean Islands. Strong southerly gale winds 9Bf over all seas. First snowfall of the year was reported in mainland mountains (Fig. 77).

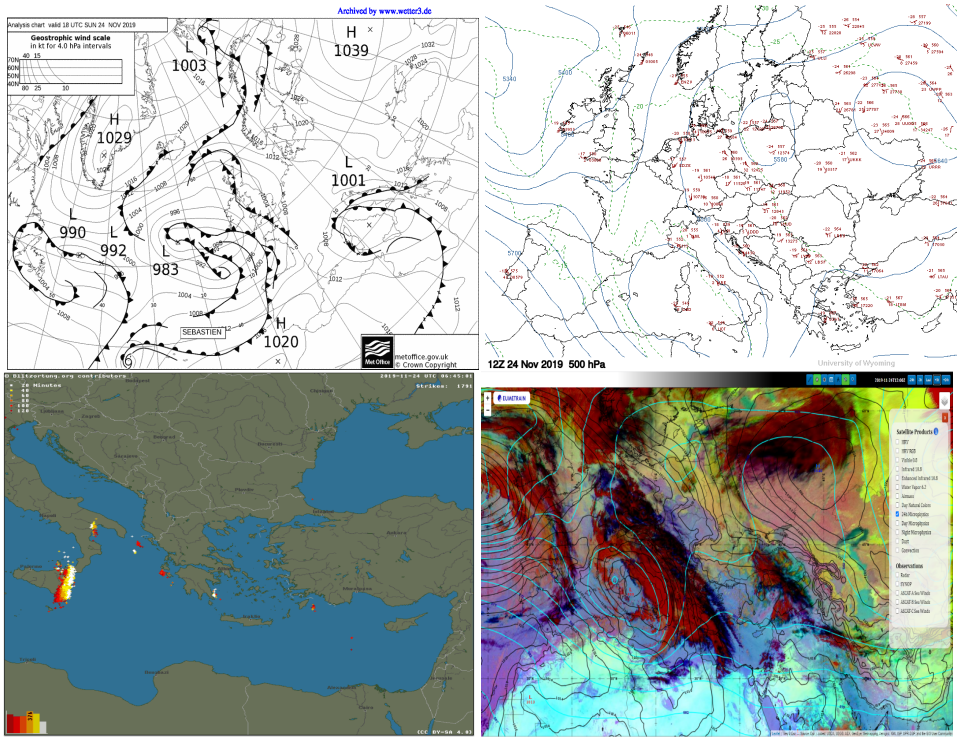


Figure 77: Upper row: MSLP maps during the event. Lower row: accumulated observed lightnings (left) and RBG air masses.

Evaluation of LPI Forecasts – Dichotomic Approach

In this section, the statistical results of the evaluation of the upscaled forecast and observation fields are presented. The methodology presented in section 3 was applied for all three test cases and the relevant plots for POD, FAR, ETS and FBI.

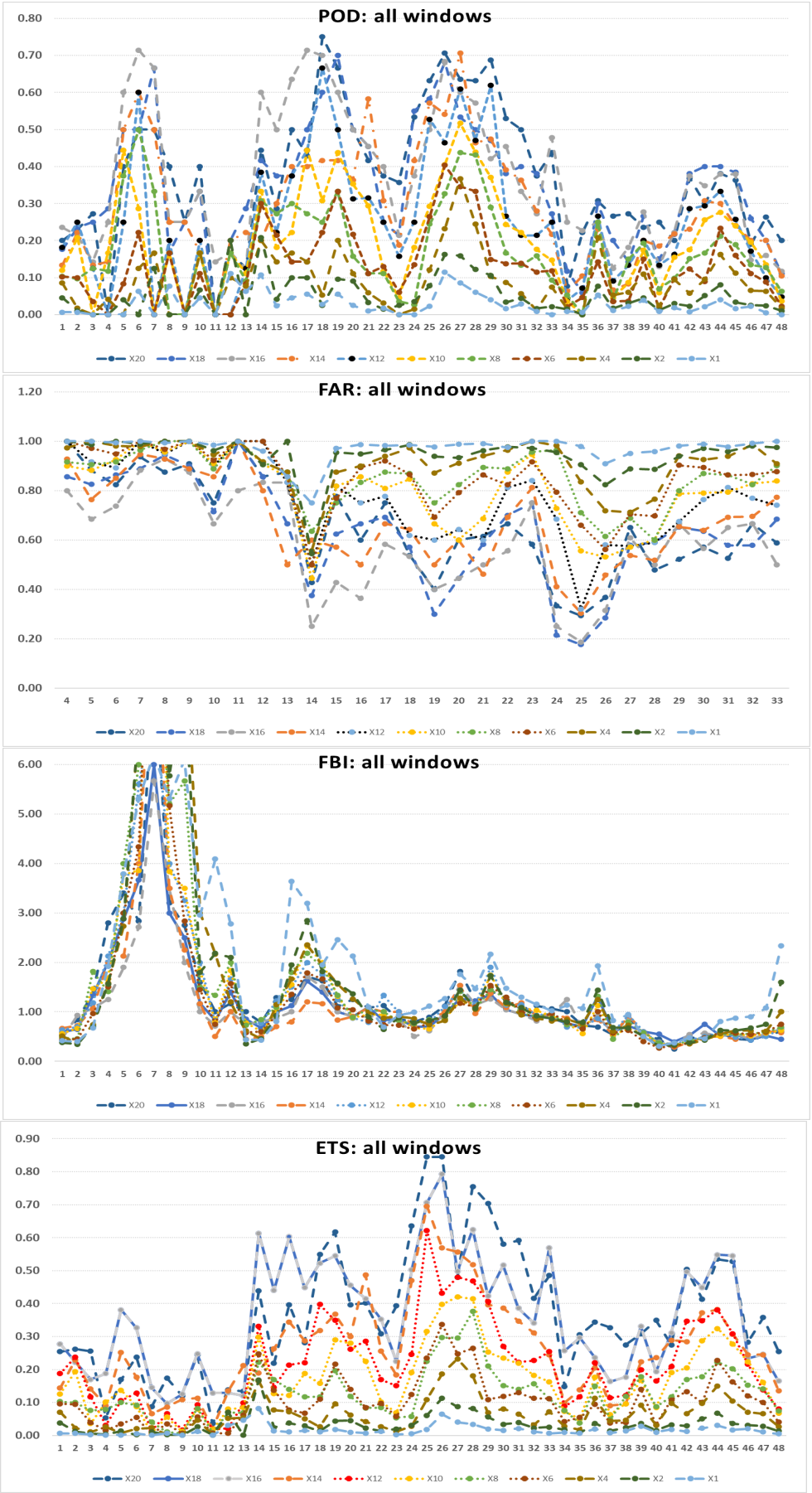


Figure 78: Test case 1 - From top to bottom: POD, FAR, FBI, ETS for various time intervals during the event and for increasing spatial resolution.

Remarks from Test Case 1: No skill for LPI forecasts for the first 12h of the event

POD: reduced skill during afternoon hours, improved performance for scales larger than 16x0.04~64km

FAR: Improved performance for scales higher than 14x0.04~56km, no variation in performance with lead time.

FBI: no impact of the upscaling approach in the performance, high overestimation in first 10 forecast hours

ETS: performance does not increase linearly with increased resolution, optimum skill in most time intervals when the 64km resolution is applied.

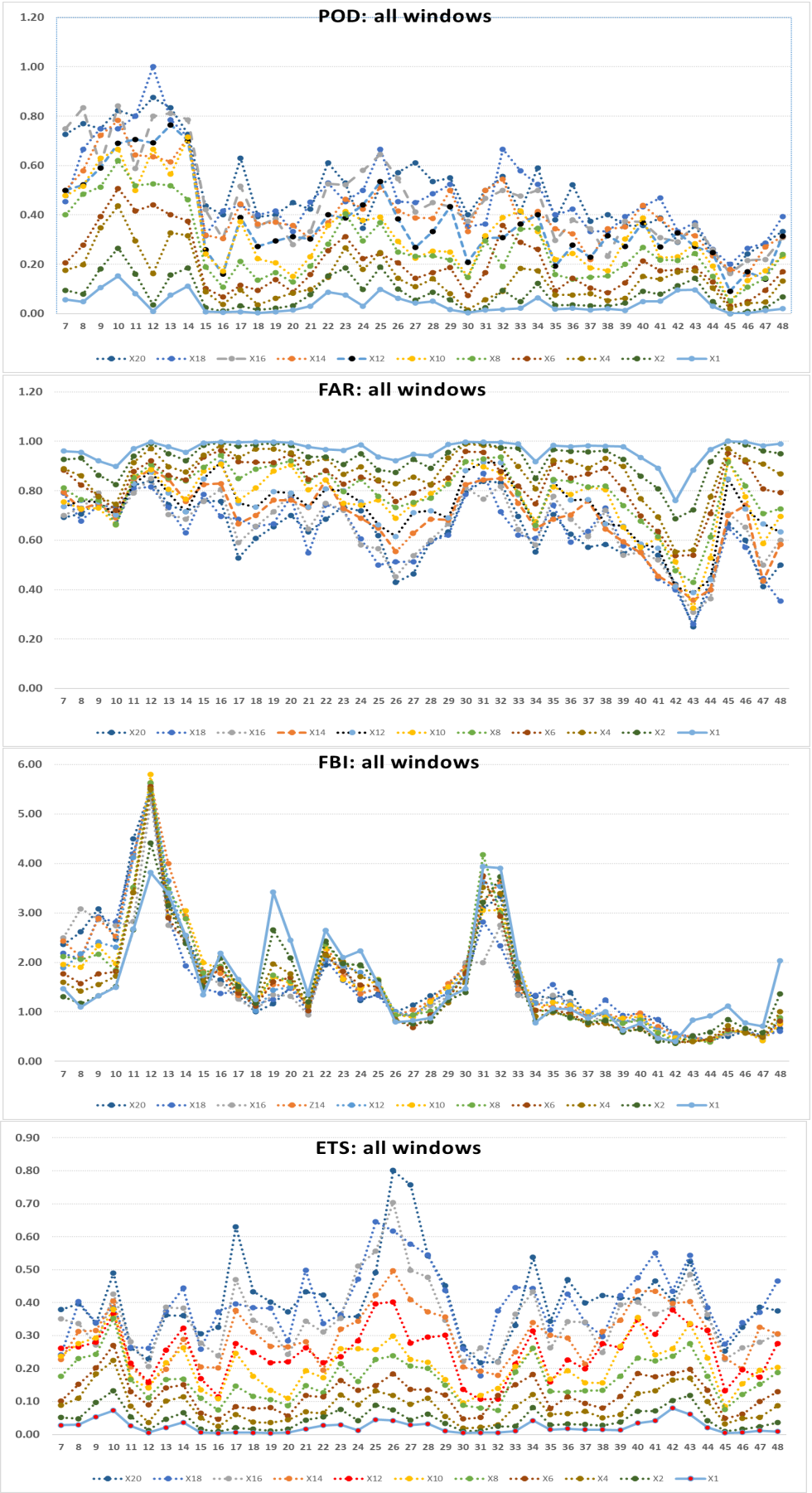


Figure 79: Test case 2 - From top to bottom: POD, FAR, FBI, ETS for various time intervals during the event and for increasing spatial resolution.

Remarks from Test Case 2: No skill of forecasted fields in the original resolution

POD: no change in skill with lead time. Improved performance for scales higher than 12x0.04~48km with no clear improvement in further upscaled fields.

FAR: Improved skill for almost all fields during evening hours

FBI: no impact of the upscaling approach in the performance, higher overestimation for the original resolution but also for almost all upscaled forecasted fields.

ETS: performance increases linearly with window size until 14x0.04deg~56km

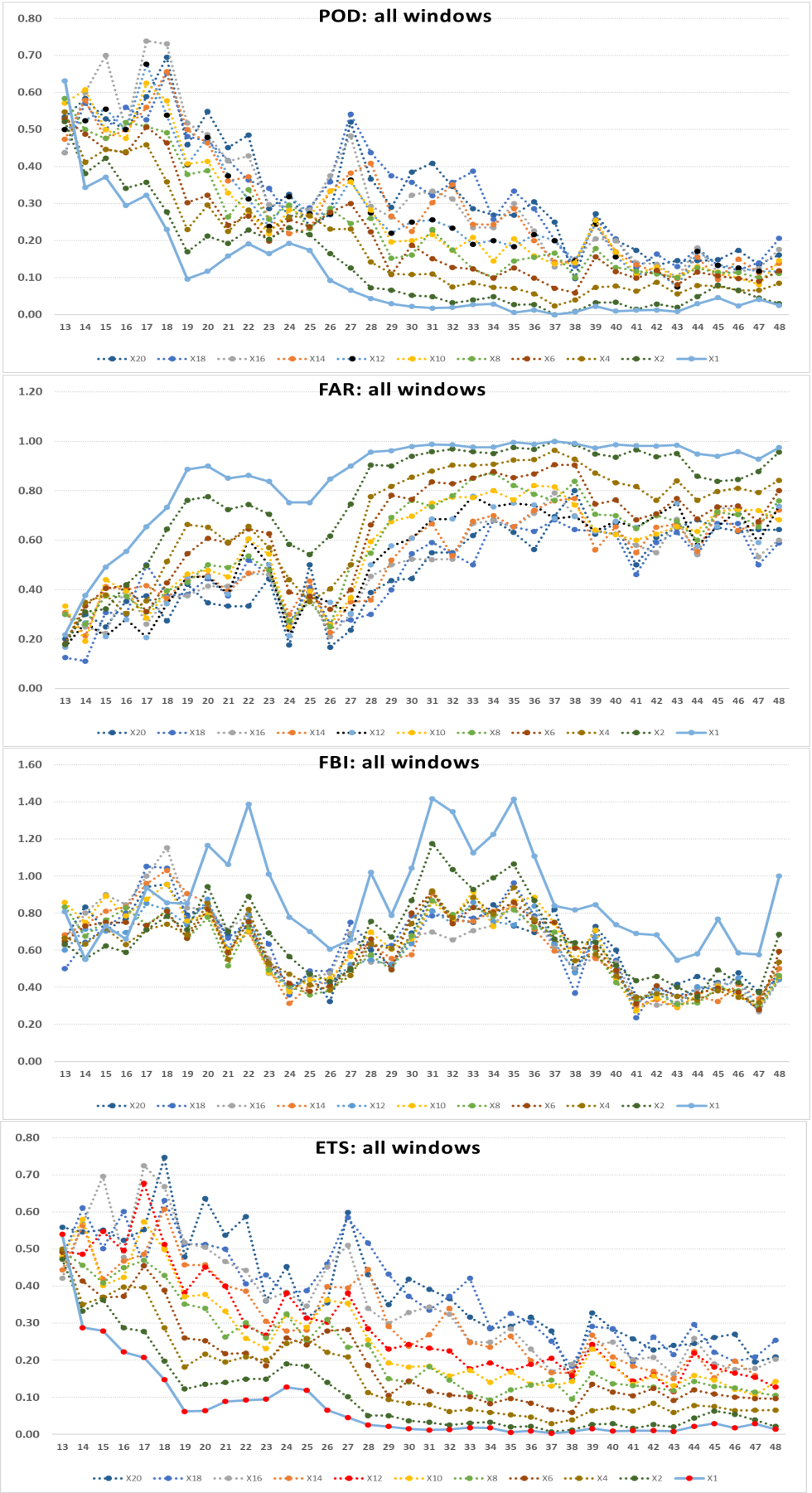


Figure 80: Test case 3 - From top to bottom: POD, FAR, FBI, ETS for various time intervals during the event and for increasing spatial resolution.

Remarks from Test Case 3: Good performance of LPI forecasts for this event compared to the previous cases even with the original resolution.

POD: Skill is reduced with lead time

FAR: For resolution higher than 10x0.04~40km has reached already adequate skill.

FBI: Small underestimation of LPI predictions is shown in all upscaled grids

ETS: Performance increases linearly with window size. For windows higher than 40km there is a good skill in LPI forecasts.

Evaluation of LPI Forecasts – SAL Approach

During the application of Structure, Amplitude, Location (SAL) spatial methodology the original resolution of both forecast and observed fields was used.

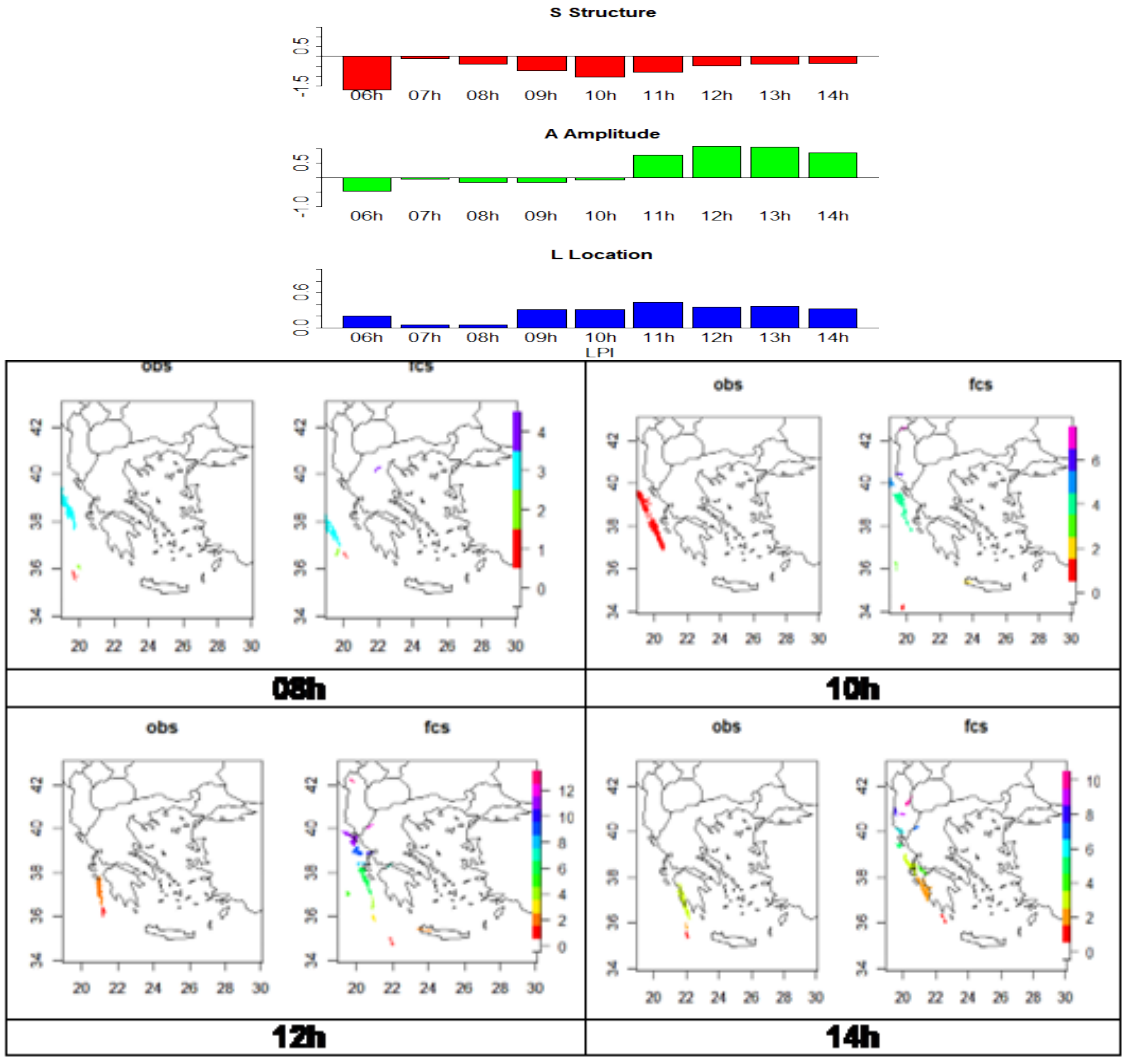


Figure 81: Test case 2 – Objects matching for several time windows during the event. SAL components with respect to forecast horizon.

The S values are negative, indicating that the model predicts sharper objects than the ones observed.

The A is positive with value higher than 0.5 during afternoon hours (total LPI overestimated as shown in FBI index in upscaling approach).

The L parameter is also increases after 09h, indicating some differences in the location of objects with respect to the observed ones.

Remarks from Test Case 3:

The S values trend is variable with horizon time and seems that model predicts more widespread objects in the beginning and around the end of the forecasted period.

The A absolute values are smaller than 0.5 while the total LPI is satisfactorily predicted (slightly overforecasted mainly around 20-23h).

The L parameter is low (around 0.2) and shows good agreement on the location of objects with respect to the observed ones.

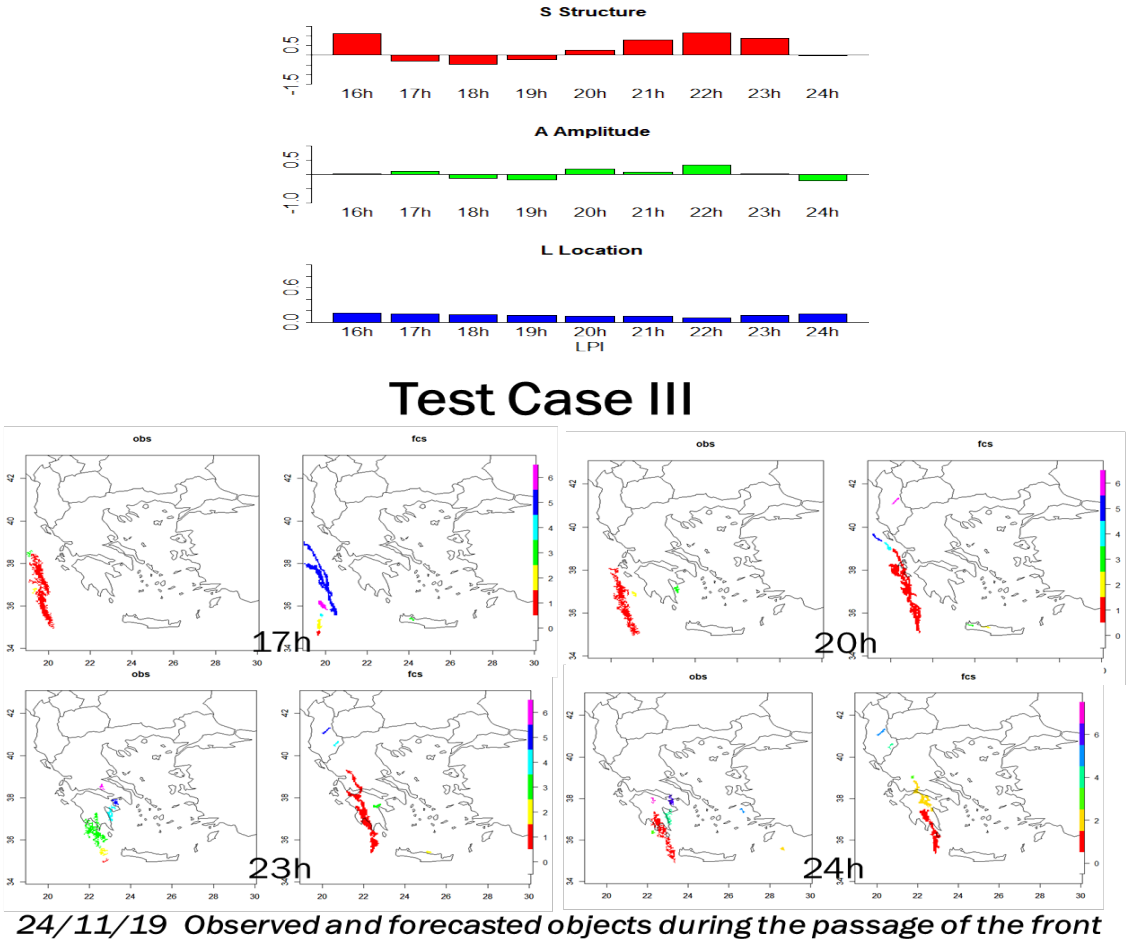


Figure 82: Test case 3 – Objects matching for several time windows during the event. SAL components with respect to forecast horizon.

Post Processed Thermodynamical Indices

Using the necessary forecasted fields in the original resolution, several thermodynamical indices were calculated and plotted according to the information provided in paragraph 4. Appropriate colour pallets were utilized for each index in order to notify areas with high possibility of presence of convection.

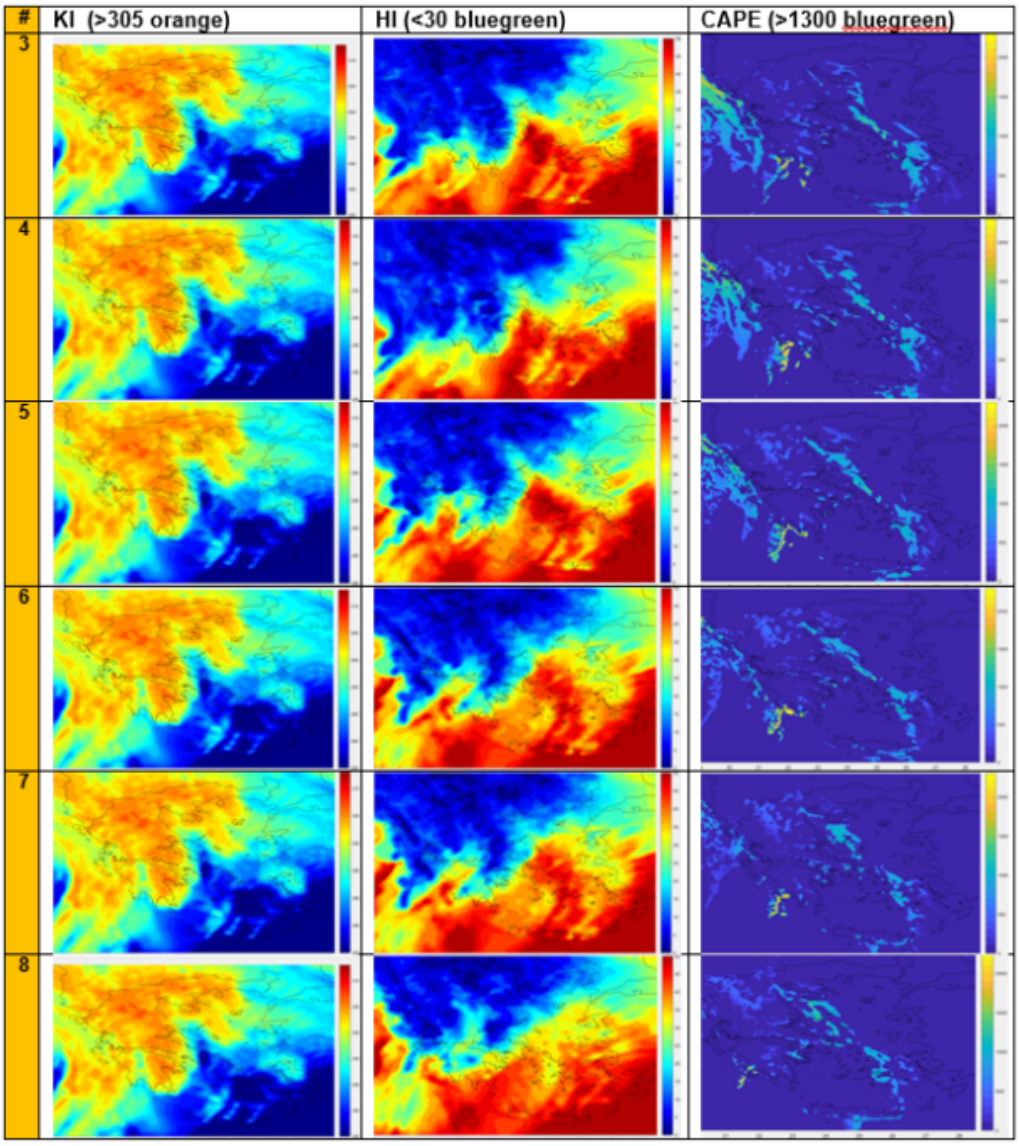


Figure 83: Test case 1 – Presentation of various thermodynamical indices during the evolution of the event with indication of color/threshold that corresponds to high convection probability.

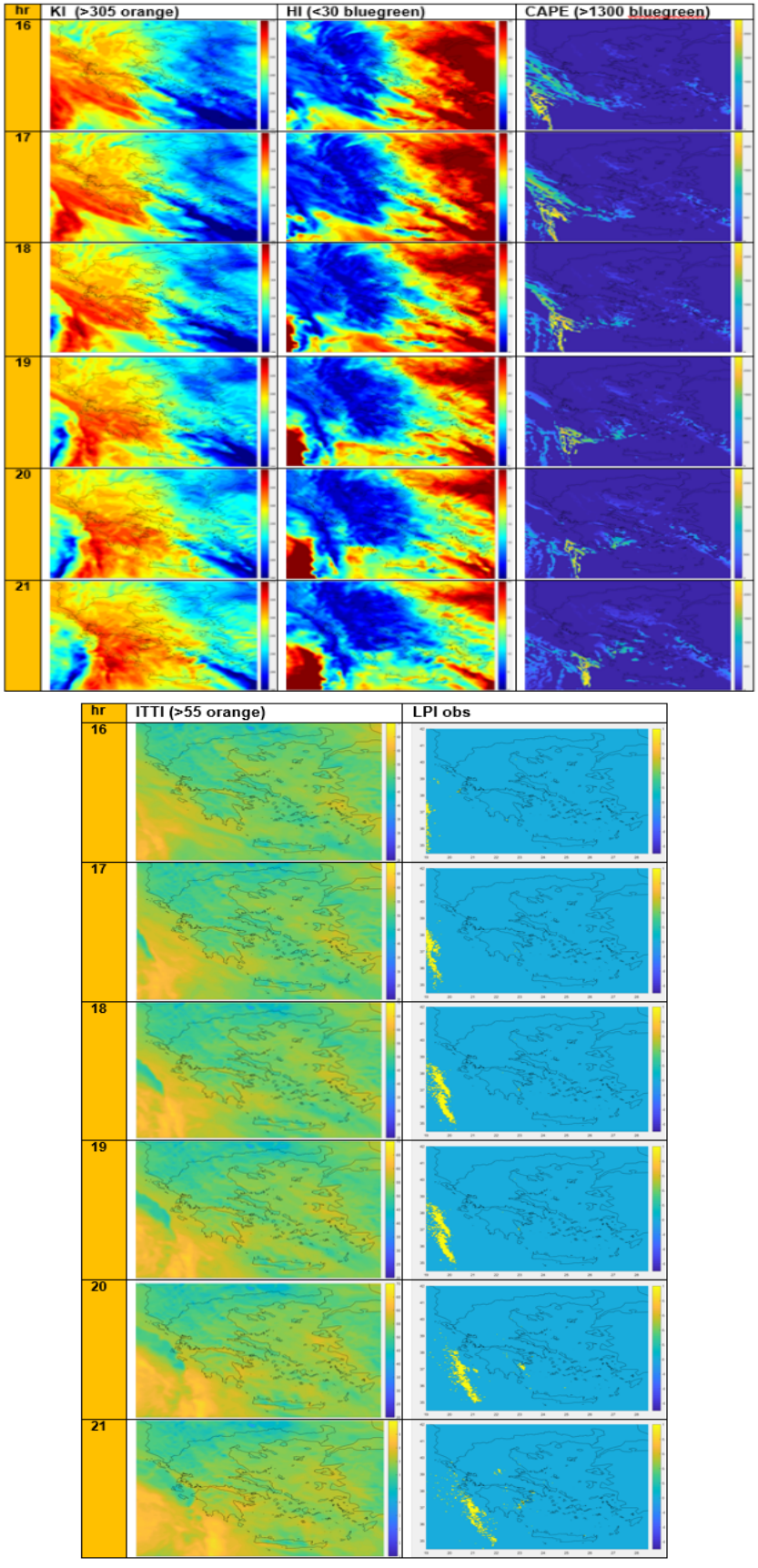


Figure 84: Test case 2 – Presentation of various thermodynamical indices during the evolution of the event with indication of colour/threshold that corresponds to high convection probability.

Recommendations

The main outcomes from the work performed for Task 3.5, can be summarized as follows:

1. It is necessary to derive upscaled LPI products in resolution larger than 40km (10times the original one), in order to gain reliability in the forecasts. From the analyzed events, the performance of COSMOGR4 for LPI seemed to be strongly dependent on weather regimes.
2. LPI raw values need to be thresholded according to the area and period examined. Further study for longer periods is necessary in order to determine what thresholds are appropriate for the specific geographic area.
3. Thresholds for thermodynamical indices associated to severe thunderstorms need to be appropriately defined to provide useful indication of a thunderstorm area. Default values often do not apply.
4. The lightning potential index produces reliable lightning information during stronger storms, much like observed in observational data. A general overestimation of the presence of lightning was derived when native resolution was used.
5. 'Forecasters would be able to anticipate lightning activity from other model outputs such as CAPE or postprocessed thermodynamical indices even with less accuracy in the position, For forecasters the added value of direct LPI forecasts used proved to be very small, or not present at all.
6. 'Probably, the LPI is somewhat better at distinguishing lightning-producing storms and this may be of importance to some user groups.

References

1. Yair Y, Lynn B, Price C, Kotroni V, Lagouvardos K, Morin E, Mugnai A, Del Carmen Llasat M. Predicting the potential for lightning activity in Mediterranean storms based on the Weather Research and Forecasting (WRF) model dynamic and microphysical fields. Journal of Geophysical Research Atmospheres. 2010;115 :1–13.
2. Miller RC (1967) Notes on analysis and severe storm forecasting procedures of the Military Weather Warning Center: Tech. Report 200, AWS, USAF.
3. *Johnson DL (1982). A stability analysis of AVE-4 severe weather soundings. NASA TP-2045 13: 8.*
4. *Grieser, J., 2012. Convection parameters. Selbstverl.*
5. *Saunders, C. P. R. (2008), Charge separation mechanisms in clouds, Space Sci. Rev., 137, 335– 353.*
6. *Moncrieff, M. W. and Miller, M. J. (1976): The dynamics and simulation of tropical cumulus and squall lines; Quart. J. Roy. Meteor. Soc., 102 373-394.*

6.6 Comparative verification of NWC and NWP results using spatial verification methods as part of the SINFONY project at DWD

Gregor Pante and Michael Hoff, DWD

Introduction

Germany is exposed to various kinds of high impact weather phenomena. Strong impacts are expected from convective events during summer which happen to be especially hard to predict. The Seamless Integrated Forecasting System (SINFONY) project at DWD focuses on such events, which mostly take place on the kilometre scale. One aim of the project is therefore the development, adaptation, and operationalization of innovative, spatially based verification methods of the entire process chain of the integrated forecasting system consisting of data assimilation, nowcasting and numerical short-term prediction. The advantage of spatially based verification methods is that exact matching of forecasts and observation no longer needs to prevail to obtain good scores because these methods circumvent the "double penalty" problem, i.e. a miss due to a displaced observation event and a false alarm due to a displaced forecast event. Following Gilleland et al. (2009) there exist mainly four categories of spatial verification – neighbourhood (or fuzzy) and scale-separation basically applying filtering methods, as well as feature (or object) based and field deformation basically yielding information about displacements. In the SINFONY project, we decided to apply neighbourhood as well as object-based verification methods. Both methods are well established and cover a huge amount of information which is helpful for model development, user interpretation and many more.

The neighbourhood (also known as fuzzy) approaches compare values of forecasts and observations in space–time neighbourhoods relative to a point in the observation field. Properties of the fields within neighbourhoods (e.g., mean, maximum, existence of one or more points exceeding a certain threshold) are then compared using various statistical summaries, which are often simply the traditional verification statistics. Such comparisons are typically done for incrementally larger neighbourhoods so that it is possible to determine the scale at which a desired level of skill is attained by the forecast (Gilleland et al., 2009). The neighbourhood methods apply a smoothed filter on the original field(s). Summary statistics, such as traditional verification statistics, can be applied to the smoothed field. The process is typically repeated using increasingly larger neighbourhoods. The most established neighbourhood method is called Fractions-Skill-Score developed by Roberts and Lean (2008).

Of particular interest, especially in SINFONY, are object-based methods which require a threshold-linked object identification algorithm. It is applied to pixel-based forecast and observation fields of radar reflectivity. The resulting objects contain certain attributes regarding their geometry (e.g., centroid, area), intensity (e.g., min, max), and forecast information (e.g., trajectory). In SINFONY, we focus on the object-based evaluation metric called median of maximum interest (MMI) after Davis et al. (2009) to assess the quality of the predicted precipitation objects. The object-based evaluation is extended to cope with ensemble forecasts. Besides basic single member verification a new technique to define a so called "pseudomember" (Johnson et al., 2020, J20 hereafter) is analyzed. The pseudomember comprises a reasonable and representative selection of objects from all ensemble members that have locally the highest probability of occurrence.

Data and methodology

Data sets

In SINFONY will be a variety of data available for verification. The case study period of the underlying data will be mentioned in the respective section.

Grid-based data

For numerical weather prediction, NWP, the underlying model is the regional ICON-D2-EPS in a quasi-operational setup since 2019. Before 2019, we were using COSMO-DE-EPS in a quasi-operational setup. The EMVORADO operator (Zeng et al., 2016) simulates synthetic radar reflectivity for each of the 17 polarimetric Doppler-C-Band radar systems in the DWD radar composite. Subsequently, the model volume scans will be processed by POLARA and mapped onto a Radolan grid with a horizontal resolution of 1km. We are using 40 members for data assimilation and 20 members for the forecast of up to 8 hours.

For nowcasting, we are using STEPS DWD with a localization filtering approach (planned to submit in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing) to generate an ensemble with 30 member (20 member used for verification) with a nowcasting time of two (or four) hours. All nowcasting data are on the 1km horizontal resolution Radolan grid.

Object-based data

For object identification, we are using our in-house product KONRAD3D. With the help of adaptive thresh-olding schemes and other filtering methods, which will not be specified in this report, KONRAD3D identifies cell objects in each radar volume scan. The default basic threshold for object identification is 35dBz whereas a subcell within such regions must fulfill a minimum-maximum difference to the basic threshold of at least 7dBz. This means that cells mostly obtain a minimum value of 42dBz. By optimized combination of objects in each radar volume scan, three-dimensional objects will be built taking into account the entire DWD radar network.

KONRAD3D is used for object nowcasting as well. Currently in development is an ensemble based object nowcasting which will unfortunately not be available for the current study. However, since EMVORADO simulates reflectivity for all radars and respective volume scans, KONRAD3D can be applied to the NWP forecasts, described in the previous section. Therefore, we can fall back on an 20 member ensemble object-based NWP forecast of 8 hours with temporal resolution of five minutes to test our object-based verification methods. Further, a comparison of 1-moment vs. 2-moment microphysics scheme will be made. As the latter is able to produce higher reflectivity, it is expected to better capture extreme events.

Combined product Seamless combination of nowcast and model forecast, grid-based and object-based.

Spatial verification methods

Neighbourhood-based methods

We apply mostly well-known neighbourhood-based verification methods to our data. The most established method is the fractions skill score, FSS, (NO-NF) by Roberts and Lean (2008). Further, we implemented the minimum coverage method (NO-NF), Fuzzy-logic (NO-NF), fuzzy-logic with joint probabilities (NO-NF), multi-event contingency table (SO-NF) and pragmatic approach (SO-NF) for which the reader is referred to as Ebert (2008). All necessary information about the underlying methods can be found in this publication. Since the above mentioned methods for building a contingency table from neighbourhood

probabilities have weaknesses in their bias behaviour, we implemented the neighbourhood-based contingency table including errors compensation by Stein and Stoop (2018). This method uses a practical approach in which it the same number of misses and false alarms in a certain neighbourhood compensate each other to hits and correct negatives, i.e. it is a correct forecast in the respective neighbourhood. A positive side effect of this method is that the frequency bias is independent of the neighbourhood size (small deviations on domain edges are possible), which makes it quite practical using it for verification.

Another useful method we implemented is the displacement estimation of precipitation fields based on fractions skill score by Skok and Roberts (2018). The authors used the FSS = 0.5 threshold for a useful forecast to estimate a global distance metric. The results are quite promising even though the method is not applicable for frequency biases larger than two and lower then 0.5. Also for frequency biases larger than 1.5 (< 0.75) the method exhibits shortcomings. However, for the remaining data, the displacement metric is a useful information apart from the classical categorical verification metrics. To go one step further, G. Skok presented a new metric called displacement from NSS (neighbourhood skill score) at 2020 International Verification Methods Workshop. Further, Skok showed that the results are closer to the real displacement and also in more realistic cases the score showed more reliable results. The Displacement-NSS is no more limited to small biases which makes it quite useful for application in our SINFONY project. Therefore, with the help of G. Skok, we implemented this metric as well. However, the deviation of the NSS displacement from real displacement becomes larger the closer precipitation objects are to the domain edges. Up to now, we did not correct this fact in our verification analyses but postpone it to future work.

Another useful method, we implemented, is Neighbourhood-Ensemble-Probabilities (NEP) proposed by Schwartz et al. (2010). Here, the thresholding, neighbourhood-smoothing (for different box lengths) will be done for all M ensemble member separately. Finally, the resulting M neighbourhood probabilities will be averaged to obtain NEP. On the NEP field, all above described methods can be applied, however, not all methods will give benefits for using NEP. The most reliable method in combination with NEP is FSS. The NEP is most beneficial for smaller neighbourhood sizes around a certain point of interest. For larger neighbourhoods, the effect will be smoothed out or the areas of precipitation probabilities become to large in comparison with the observation.

Finally, we implemented reliability and ROC diagrams for analysing our grid-based deterministic and ensemble data. As reference, however, we made a compromise and took only binary observation into account, since otherwise the huge quantity of verification data is not manageable in an operational framework.Further implementations are planned for the future.

Object-based methods

Total Interest and Median of Maximum Interest

The TI (Davis et al., 2009) is a measure for the similarity of two objects with respect to the objects' attributes. For each selected attribute i of an object pair j a "fuzzy logic function" (F) is defined in order to transform the value of i into the interval [0,1]. For example, the function of the centroid distance (FCD) – one attribute of an object pair – is defined to be equal 1 if CD is less than 10 km, then linearly decreases with increasing CD and equals 0 for CD larger than 100 km. The F-values of different attributes result in the "interest" (I) by multiplying with weights (w) and confidence factors (c). The TI of an object pair j, finally, is the weighted sum of all I-values of all considered attributes i (Davis et al., 2009):

$$TI_j = \frac{\sum_i I_{ij}}{\sum_i w_i c_i}$$
$$I_{ij} = F_{ij} w_i c_i$$

Attribute w,	%	c	fmin	fmax
Centroid distance	28	Area ratio	10 km	100 km
Minimum boundary distance	40	1	5 km	50 km
Area ratio	19	1	0.0	0.8
Intersection area ratio	13	1	0.0	0.25

Table 17: Attributes and parameters used to calculate the total interest TI. fmin and fmax are the lower and upper limits below and above which the fuzzy logic function of the respective attribute takes its minimum, respectively maximum value.

In the presented analysis we employ the settings as described in Davis et al. (2009) and listed in Table 17 to calculate the TI. Having one set of observed and one set of predicted objects, the TI-values of all possible object pairs are calculated. They fill the so called TI matrix which contains all observed objects as columns and all predicted objects as rows. The next step selects the maximum values along each row (column) and adds them as a new column (row) at the right (bottom) of the TI matrix. The median over all these maximum values builds the final score for the object-based ensemble verification, i.e., the median of maximum interest MMI.

Ensemble forecasts

The object-based evaluation of ensemble forecasts is one major challenge in the verification for two reasons. First, the amount of objects to be processed can be very large depending on the weather situation and number of ensemble members. And second, new methods must be developed to reveal a fair score. Two rather simple ideas are the verification of the objects from each single ensemble member separately or of the merged set of all objects from all members. The first one yields simply the quality of each member and can additionally provide information about the spread of the ensemble. The second one is very likely to generate so called "over-forecasting", i.e., the combined set of objects comprises much more objects than the observation which may generate many false alarms. Therefore, a third method is analyzed in which a reasonable selection of objects is chosen to build the so called "pseudomember" which comprises the objects from all ensemble members that are locally the most representative ones of the ensemble distribution (J20).

The selection of objects for the pseudomember follows five steps as described in J20:

1. "Make a list of all objects in the forecast ensemble, together with the objects' probabilities, calculated from the percentage of ensemble members with a matching (i.e., total interest > 0.2) object.

2. Sort all of the objects by probability, breaking ties according to the average total interest with all the objects from other ensemble members that it matched to.

3. Add the highest probability object to the object list of the pseudomember.

4. Remove from consideration the added object, as well as all matching objects in other

members that contributed to the probability of the added object, leaving a new, smaller list of objects.

5. Repeat from step 2 until no objects remain in the list of ensemble forecast objects."

Here these steps are performed for constructing the pseudomember but another matching criterion (first step of J20) was used. For the comparison of one specific object from one member with all objects from all other members, the TI of this specific object with all other objects is calculated as:

$$TI = \frac{2 \cdot F_{CD} + F_{AR}}{3}$$

where F_{CD} and F_{AR} are the interest functions of centroid distance (CD) and area ratio (AR). These functions are defined as

$$F_{CD} = \begin{cases} 1 & CD < CD_1 \\ \frac{1}{2} \cdot \left[\cos\left(\frac{CD - CD_1}{CD_2 - CD_1} \cdot \pi\right) + 1\right] & CD_1 \leq CD \leq CD_2 \\ 0 & CD > CD_2 \end{cases}$$

$$F_{AR} = \begin{cases} 0 & AR < AR_1 \\ \frac{1}{2} \cdot \left[\sin\left(\frac{AR - AR_1}{AR_2 - AR_1} \cdot \pi - \frac{\pi}{2}\right) + 1\right] & AR_1 \leq AR \leq AR_2 \\ 1 & AR > AR_2 \end{cases}$$

Below CD1 and AR1 and above CD2 and AR2, which are set to CD1 = 10 km, CD2 = 70 km, AR1 = 0, and AR2 = 0.8, the interest functions take their minimum (0), respectively, maximum (1) values. For object pairs to be defined a match the TI must exceed a value of 0.7. This criterion limits the ranges of CD and AR within which matches are possible. Hence, if CD is larger than 38 km no match can occur even if AR was perfect while no matches occur for AR below 0.16 even if CD was perfect.

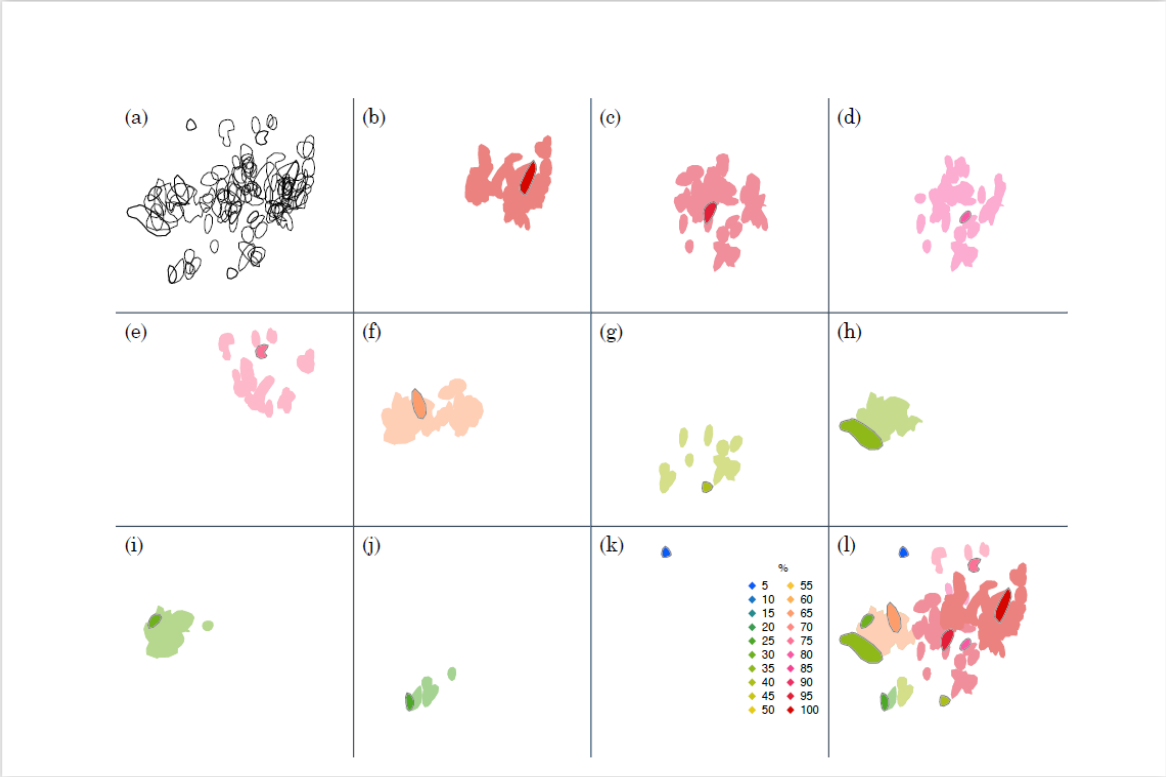


Figure 85: Procedure of selecting the objects of the pseudomember for a forecast initialized on 30 May 2016 12 UTC with a lead time of 3 hours. All objects of all ensemble members in a given region are shown in panel (a). Panels (b)–(k) depict the single pseudomember objects (gray bordered polygons) according to their probability of occurrence (colours). The lighter colours around these objects mark the uncertainty regions, i.e., the unified area of all objects from other members that were defined a "match" with the respective pseudomember object. The combined result with all pseudomember objects is given in panel (l). Coloured areas of the uncertainty regions are stacked on top of each other with increasing probability, hence, regions with low probability can be covered by those with higher probabilities.

Figure 85 illustrates the procedure of selecting the pseudomember objects following the steps described above. Technically, the pseudomember is a list of polygons, i.e., the selected objects, together with their probabilities of occurrence and uncertainty regions. The probability of occurrence p (colour scale in Fig. 85) is the percentage of ensemble members with at least one matching object. The member of the object in consideration itself is counted as well, hence, for a 20 member ensemble, as used in this study, p varies in 5% steps between 5% and 100%, where 5% means no other member has a matching object and 100% all other members have at least one matching object. If a member has more than one matching object all these objects are removed from further consideration (step 4 in the description above) but this member still counts as only one member with regard to the probability. The uncertainty region of a pseudomember object is the unified area covered by all the matching objects from other members (light colours in Fig. 1). In the example one object has matching objects in all other ensemble members and gets a value of p = 100% (Fig. 85b). The probability of the subsequently selected objects decreases until only one object remains which has no matching objects in other members and therefore p = 5% is assigned to this object (Fig. 85k).

Results

The different spatially based verification methods described before are applied to predictions from the SINFONY reference period between 27 May – 25 June 2016. This early summer period is characterized by almost daily strong convective activity over Germany. Unfortunately, only COSMO-DE-EPS runs are available for this time period.

Neighbourhood-based methods

In this section, we show some representative results from the SINFONY reference period in May/ June 2016.

Figure 86: FSS tiles plots for reflectivity (dBz) averaged over the SINFONY reference period (27 May – 25 June 2016) and over all initial times (11 – 15 UTC) and all ensemble members (1 – 20, incl. NEP). The top row shows results for a lead time of 30 minutes, the bottom row for 4 hours. COSMO-DE-EPS 1-moment microphysics scheme (left panels), 2-moment microphysics scheme (middle panels) and STEPS nowcasting (right panels). Greenish colours represented a skilful FSS (≥ 0.5), reddish colours represent non-skilful FSS (< 0.5).

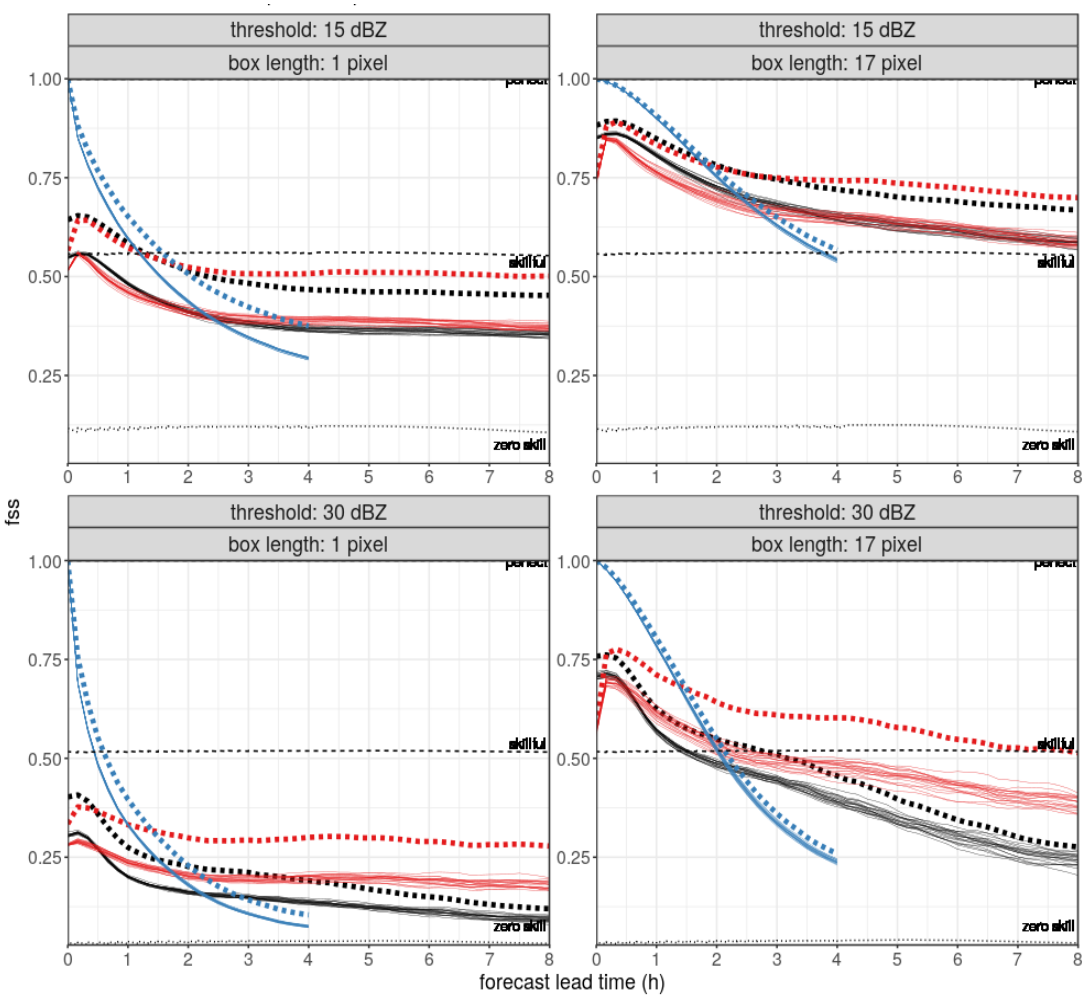


Figure 87: FSS as a function of lead time for reflectivity (dBz) averaged over the SINFONY reference period (27 May – 25 June 2016) and over all initial times (11 – 15 UTC). The top row shows results a threshold of 15dBz, the bottom row for 30dBz. The left column shows results for a box length of 1pixel (1km), i.e. no neighbourhood, the right columns for 17pixel (17km). Thin solid lines show the FSS of all ensemble members, the thick dashed line shows the FSS of the NEP field. In black 1-moment microphysics scheme NWP, in red 2-moment microphysics scheme NWP and in blue STEPS-DWD nowcasting.

A first overview of the quality of the forecasts is given by Fig. 86. It shows FSS tiles plots for 30 minutes lead time (upper row) and 4 hours lead time (bottom row), as well as three different model setups, COSMO-DE-EPS 1-moment microphysics scheme (left panels), 2-moment microphysics scheme (middle panels) and STEPS nowcasting (right panels).

Aggregated over all parameters, the FSS shows normal behaviour, i.e. increasing values with increasing box length (neighbourhood size) and decreasing values with increasing thresholds. The STEPS nowcasting (right panels) is, as expected, of better quality after 30 minutes in comparison to the NWP setups. Especially the higher thresholds show better scores in the nowcasting, mostly because the NWP is not able to produce such high reflectivity. However, after 4 hours lead time (lower panels), the NWP quality is superior to nowcasting quality, which is not surprising since the nowcasting does not include dynamical information.

Fig. 87 shows the FSS results for reflectivity (dBz) as a function of lead time, aggregated of the SINFONY reference period and all initial times (11 – 15 UTC). The top row shows results a threshold of 15dBz, the bottom row for 30dBz. The left column shows results for a box length of 1pixel (1km), i.e. no neighbourhood, the right columns for 17pixel (17km). Thin solid lines show the FSS of all ensemble members, the thick dashed line shows the FSS of the NEP field. In black 1-moment microphysics scheme NWP, in red 2-moment microphysics scheme NWP and in blue STEPS-DWD nowcasting.

It can be seen that the NWP (red, black) exhibits a short spin-up phase, whereas the spin-up effect is much stronger for the 2-moment microphysics scheme (red). The reason for this was that the model produced way to many reflectivity features in the early lead times. This effect is correct for ICON-D2-EPS in 2020 and 2021 (not shown). It is obvious that the NEP (thick dashed lines) has a quite positive impact on the score, especially for smaller box lengths. This fact answers the question whether we need an ensemble for our forecasting systems.

Another powerful tool in neighbourhood verification is a respective reliability and ROC diagram. First, it must be clarified which type of observation should be taken into account. Since neighbourhood verification methods potentially produce a huge amount of data, we decided for a compromise and used the binary

observation as reference for the diagrams. Otherwise, the user has to decide which observation neighbourhood probability threshold he is interested in.

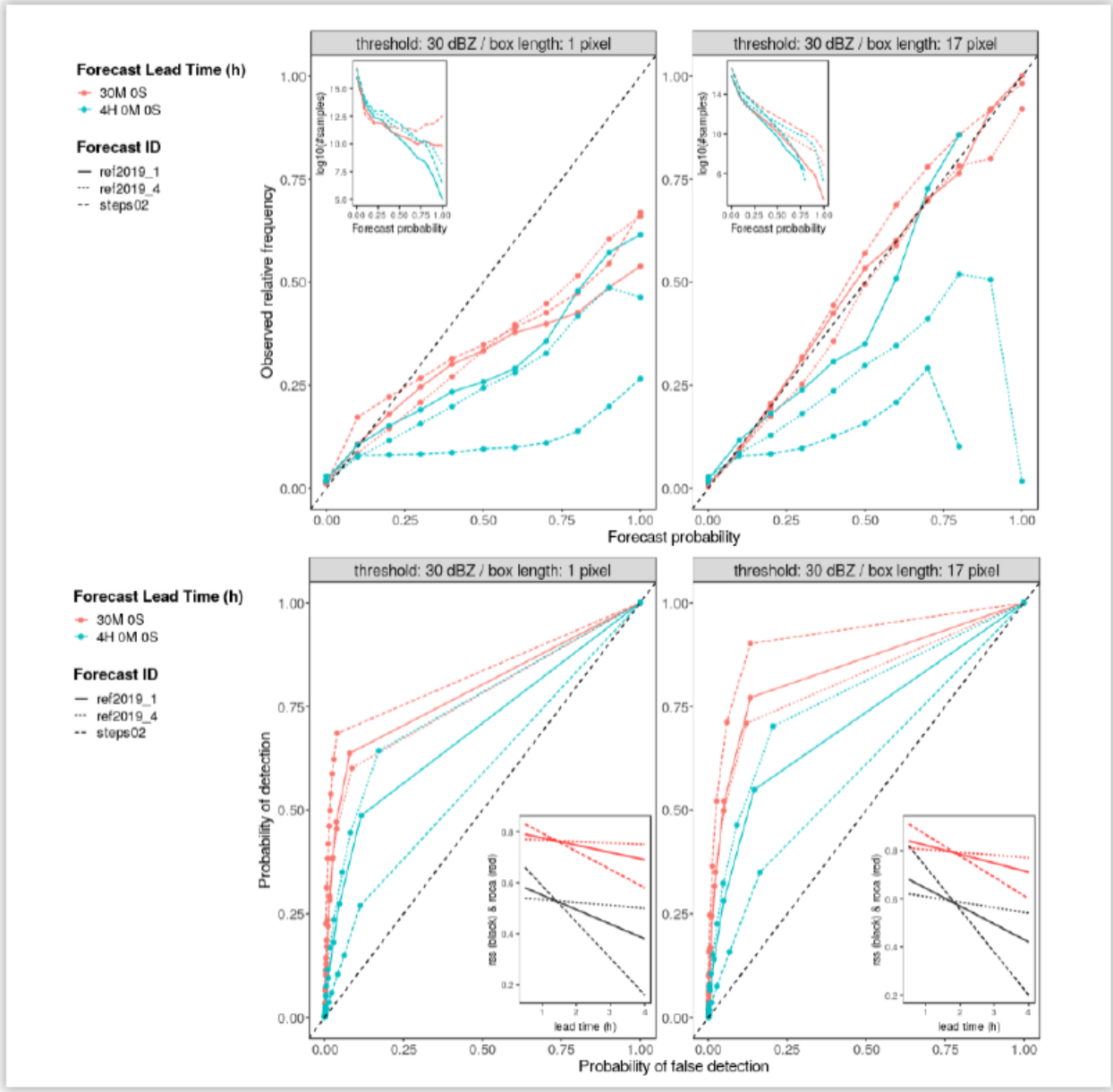


Figure 88: Reliability diagram (upper panels) and ROC diagram (lower panels) of NEP member for 30dBz and two different box lengths, pixel-based (left) and 17pixel (right). The model setups are coded as different line types. Red lines represent the lead time of 30 minutes and turquoise of 4 hours. The reference observation is of type binary.

Fig. 88 shows reliability diagrams (upper panels) of NEP member for 30dBz and two different box lengths, pixel-based (left) and 17pixel (right). The left panel shows the classical reliability diagram based on ensemble probabilities. It can be seen that there is overforecasting for almost all cases, which increases for greater lead times (4 hours, turquoise). However, when we include a neighbourhood box length of 17pixel (17km, right panel), there is almost perfect reliability of all model setups after 30 minutes lead time (red) and for some setups even after 4 hours lead time (turquoise). This confirms the fact that including a neighbourhood can exhibit a massively increased forecast quality. A similar picture is given by the ROC diagrams in the lower panels of Fig. 88. The discrimination of events

and non-events is much better when including a neighbourhood box length of 17pixel.

Another advantage of the neighbourhood-based reliability diagrams is that they can be computed even for deterministic forecasts, i.e. based on neighbourhood probabilities. This gives another great added value to forecast verification.

Finally, we want to show results for Displacement FSS and Displacement NSS developed by Skok and Roberts (2018) and Skok (2021, not yet published). In contrast to the previously described results, we have now chosen STEPS DWD nowcasting data from May/June 2021 period.

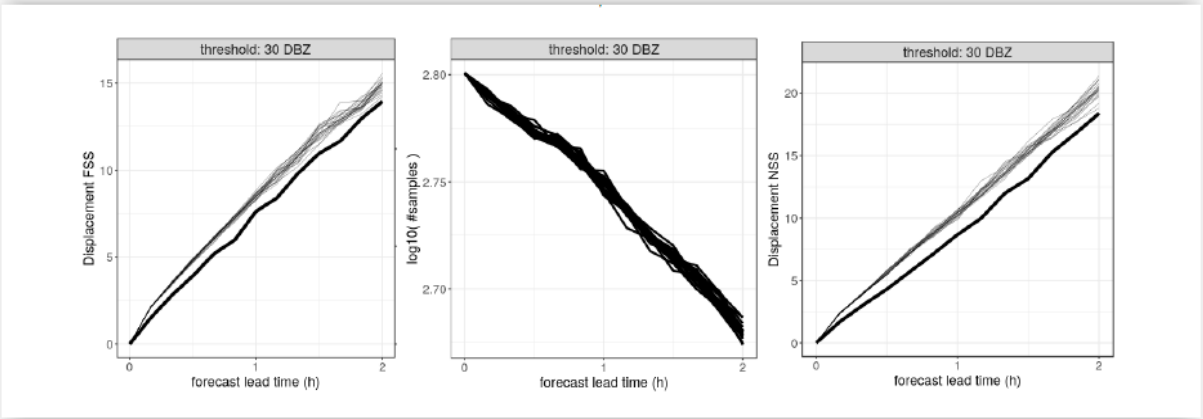


Figure 89: Displacement FSS (left), number of samples for D-FSS with $0.5 \leq \text{FBI} \leq 2$ (middle) and Displacement NSS (right) for STEPS DWD nowcasting in May/ June 2021 with 20 members. Data are aggregated over initial times from 6 – 18 UTC, 1-hourly.

The left and middle panels of Fig. 89 show the Displacement FSS (D-FSS) and respective number of samples for D-FSS with $0.5 \leq \text{FBI} \leq 2$, which are taken into account. It can be seen that the displacement is increasing almost linearly, which is in correspondence with the mechanism of nowcasting. After 2 hours of lead time, the global displacement ended up with about 14km ensemble and 13km deterministic. However, the number of samples with low bias decreased with increasing lead time.

In contrast, the Displacement NSS (D-NSS) score in the right panel of Fig. 89 has no limitation to the bias. Biased fields could simply be bias-corrected via constant factor. The displacement from D-NSS ended up at around 20km ensemble and 16km deterministic. This is slightly more than for D-FSS, however, the D-NSS score should be more confident than D-FSS. Not only because there is no bias limitation, also because some shortcomings of D-FSS are corrected in D-NSS score (see presentation of G. Skok at 8th IVMW 2020).

All in all, we found that D-FSS and D-NSS are very useful scores for interpreting other neighbourhood scores, since most of them give no information about deviations in physical parameters. Even if the absolute values are not that exact as the reality, the relative values when comparing two experiments give added value to the verification. However, there is a problem of not negligible deviations from real displacement at domain edges. Up to now, we found no solution for this but this will be done in future work.

Object-based methods

Deterministic predictions

The MMI is calculated for the nowcasting and two sets of deterministic COSMO-DE fore-

casts, the first one employing the one-moment-, the second one the two-moment microphysics scheme. Nowcasts are initialized hourly between 12 UTC and 16 UTC and run for seven hours. The model is initialized hourly between 11 UTC and 15 UTC and evaluated for the first 8 forecast hours. The shift of one hour in the initialization explains with the fact that it takes about one hour from the model start until the predictions are available. For a fair comparison the 11 UTC model forecast is therefore compared to the 12 UTC nowcast and so on.

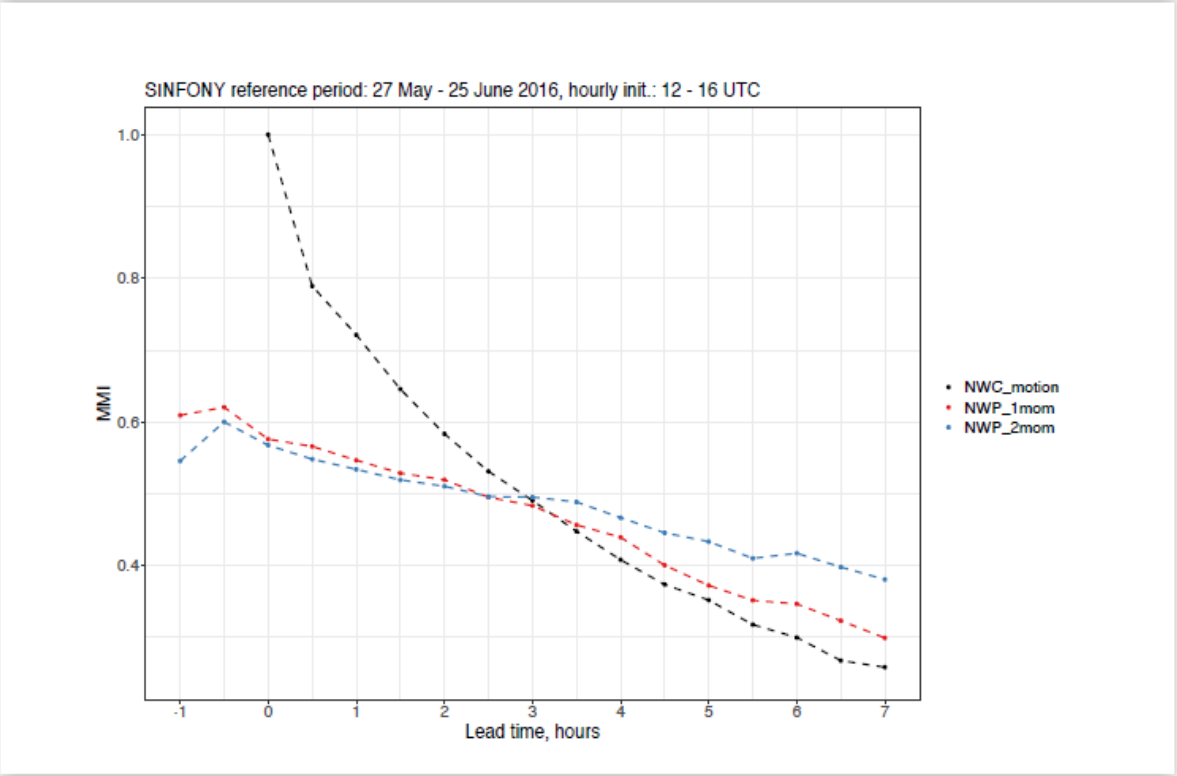


Figure 90: MMI vs lead time averaged over the SINFONY reference period (27 May – 25 June 2016) and over all initial times (12 – 16 UTC). Predictions are shown in black for the nowcasting and in red and blue for the deterministic model forecasts employing the one- and two-moment-microphysics scheme, respectively. The lead time of the model starts at -1 hour (i.e., 11 – 15 UTC), since about one hour is required for forecasts started at that time to become available.

The nowcasting starts at forecast time 0 with the perfect value of 1 (Fig. 90) because the observations serve as initialization for the nowcast and the fields are identical. The MMI decreases rapidly and is below the model forecasts after about 3 hours. The one-moment model forecasts start with higher MMI-values than the two-moment model data. At initialization, i.e., lead time -1 hour, this difference is most distinct. The artificial initialization of too many objects in the two-moment model causes the bad performance (see also discussion of Fig. 92). The MMI of the model forecasts approach after 30 minutes and the two-moment model is superior to the one-moment model after 4 hours of forecast time, i.e., 3 hours lead time in Fig. 90. From that lead time on the model forecasts perform better than the nowcasting with the clear trend that the two-moment model is superior to the one moment model.

Ensemble predictions

The analysis of ensemble forecasts is restricted to the two-moment model because its ad-

vantages at the longer lead times compared to the one-moment model.

Example of pseudomember characteristics

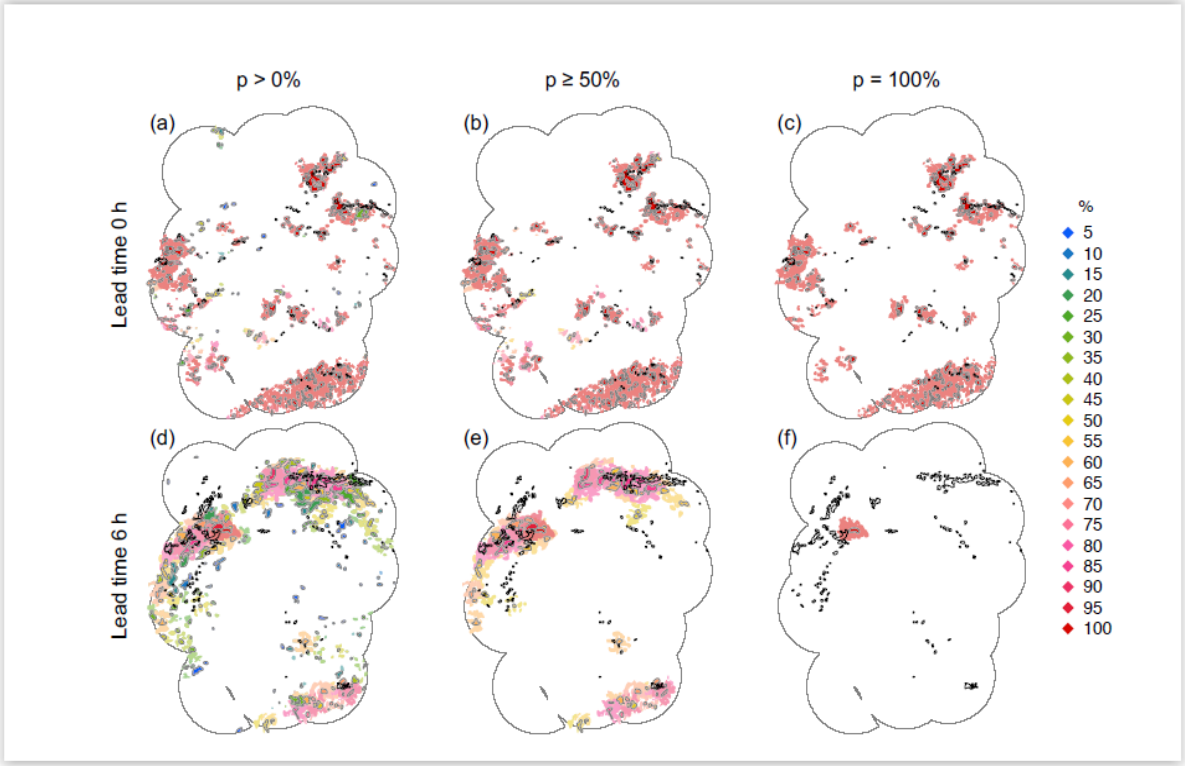


Figure 91: Observed objects (black) and objects of the pseudomember (gray-bordered) for forecasts initialized on 30 May 2016 12 UTC at lead times of 0 hours (top) and 6 hours (bottom). Pseudomember objects are coloured according to their probabilities and areas in the respective lighter colours around these objects mark their uncertainty regions (see text for further details). The effect of considering only pseudomember objects exceeding a certain probability of occurrence p is illustrated by plotting all pseudomember objects (a, d: $p > 0\%$) and only objects with $p \geq 50\%$ (b, e) and $p = 100\%$ (c, f), respectively.

Figure 91 illustrates the objects of the pseudomember, their probabilities and uncertainty regions, and the observed objects for 30 May 2016. The forecast was initialized at 12UTC. The ensemble shows little spread for a lead time of 0 hours as evidenced by the fact that most of the pseudomember objects have a probability of 100% (Fig. 91 top). For a lead time of 6 hours this has massively changed and only one object with $p = 100\%$ remains (Fig. 91f).

In comparison with the observed objects the pseudomember contains objects that represent the observations over large parts of the domain well. For lead time 0 all objects, i.e., $p > 0\%$, contain several false alarms, e.g., in the north-western and south-western part of the domain (Fig. 91a). In the South-east the pseudomember has many objects with $p = 100\%$ where several but much less objects are observed (Fig. 91c). Removing objects with low probability from consideration generally reduces the number of false alarms while introducing only few missed events over the central to western areas (Fig. 91c). This leads to a slight increase in the MMI from 0.55 ($p > 0\%$) to 0.58 ($p = 100\%$). For all p-values the number of predicted objects is clearly overestimated by a factor of 3.2 for $p > 0\%$ and still 1.7 if only objects with $p = 100\%$ are considered.

After 6 hours all objects (p > 0%) still contain false alarms over the south-western and south-eastern parts of the domain (Fig. 91d) and the total number of objects is overestimated by a factor of 1.5. Considering only objects with p ≥ 50% again removes many false alarms on the one hand but the number of missed events increases on the other hand, over the central-western areas, for example (Fig. 91e). This leads to an underestimation in the number of predicted objects, 70, compared to 137 observed objects. In comparison with lead time 0, the behaviour of the MMI is reversed. Considering all objects yields the highest MMI (0.58) although about 50% more objects are predicted than observed. Constraining the pseudomember to objects with p > 50% causes a strong reduction in the number of predicted objects leading to a lower MMI of 0.49. Constraining the objects to p = 100% is not useful for this forecast range because all but one objects have lower probabilities (Fig. 91f) yielding a MMI of 0.03.

Number of objects

The number of objects can be used as a first criterion for the quality of a forecast and it can give a rough overview about false alarms and missed events in the prediction. The mean numbers for the SINFONY reference period at all initial times between 11 and 15 UTC are shown in Fig. 92. The observations have maximum 85–100 objects at early lead times between 0 and 3 hours, i.e., 11–18 UTC depending on the initial time. The number decreases with lead time to 13 objects at +8 hours lead time, i.e., 19–23 UTC. This reflects the diurnal cycle of convective activity with most objects occurring in the afternoon that become less during the evening and early night-time hours.

The number of pseudomember objects obviously decreases with increasing values of p (Fig. 8). At lead time 0 the model has too many objects which is a well known issue in the initialization of simulations em-ploying the 2-moment microphysics scheme. These artificially initialized, unphysical objects have vanished after 30 minutes and from that lead time onward the numbers of the pseudomember objects scatter around the observed number of objects depending on p. The number of objects with p > 30% (light green in Fig. 8) represents the average number of observed objects best in both the temporal evolution with lead time and in the mean number (65 observed and 70 predicted objects).

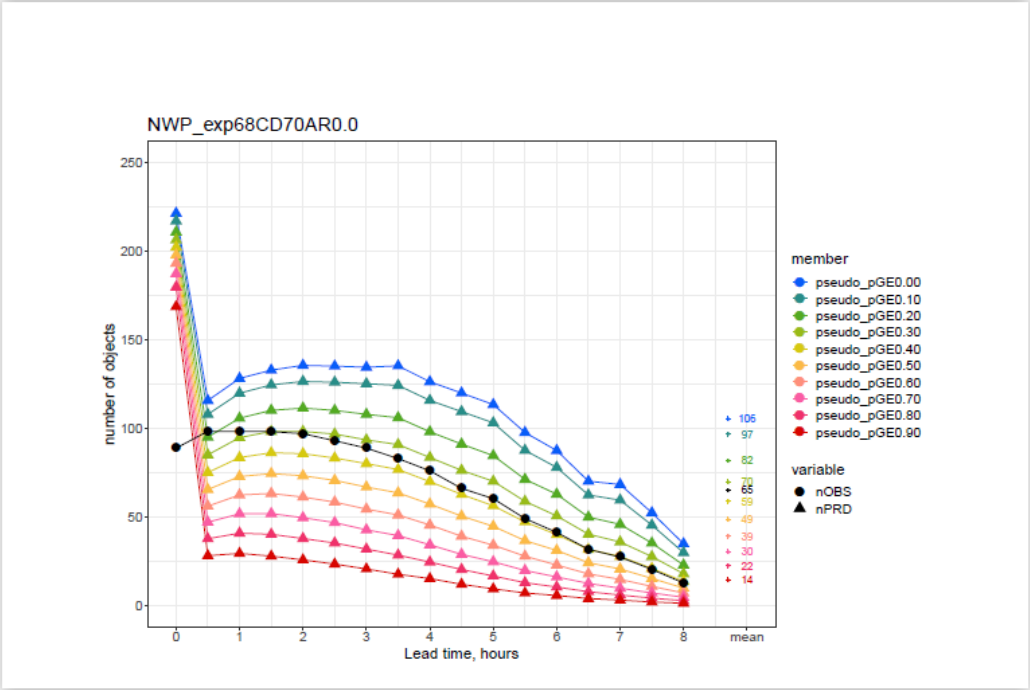


Figure 92: Number of observed (black) and predicted (colours) objects depending on the lead time, averaged over the SINFONY reference period (27 May – 25 June 2016) and over all initial times (11–15 UTC). Different colours distinguish which objects of the pseudomember are considered depending on their probability of occurrence from blue (all objects, p > 0%) to red (p > 90%). Mean values at the right are averaged over all lead times but 0 hours in order to remove the impact of artificial objects at initialization time.

MMI vs lead time

The forecast quality of different prediction types is again quantified in terms of the MMI. The following analysis comprises the MMI of the nowcasting, the deterministic forecast, all the single ensemble members, the pseudomember with p > 30%, and two "best member" selections. For the latter the MMI is calculated for each forecast and each single ensemble member separately. The best member then is selected for the evaluation. We distinguish between the best member at each forecast time step ("best member at each step") and the best member on average over forecast lead time ("best member over lead time"). For these selections the observations for all lead times are required, hence, they cannot be used as forecasts. Compared to the other real forecasts this method globally (over the entire domain) selects the best ensemble member as if one knew a priori which member will be the best for each forecast. The best member selections help to classify the quality of the other members.

The MMI of all these prediction types is illustrated in Fig. 93. The nowcasting (black) and the de-terministic forecast (dashed blue) are the same as in Fig. 90. The MMI of the nowcasting is below the different model forecasts (blue) after about 2–4 hours. The deterministic forecast is slightly better than any individual ensemble member (dotted). The quality of the pseudomember is persistently the best, except lead times -1 and 7 hours, surpassing the quality of the nowcasting after only 2 hours. The pseudomem-ber even outperforms the "best member" selections showing that it is better to do a localized selection of representative objects from the ensemble distribution than to choose the member that is globally the best. Again, the pseudomember is purely based on the ensemble forecasts while the "best member" selections need observational data for all lead times. This shows the enormous

potential of the pseudomember for the object-based forecasting of precipitation.

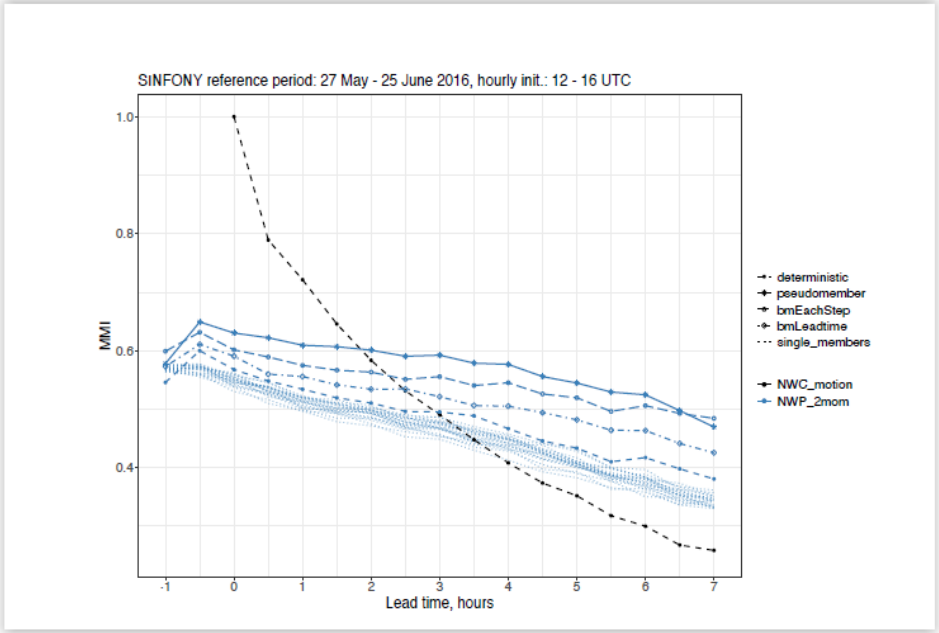


Figure 93: MMI vs lead time averaged over the SINFONY reference period (27 May – 25 June 2016) and over all initial times (12 – 16 UTC). Predictions are shown in black for the nowcasting and in blue for the model. The lead time of the model starts at -1 hour (i.e., 11 – 15 UTC), since about one hour is required for forecasts started at that time to become available. Different forecast types are distinguished by line types and symbols. The pseudomember is restricted to objects with p > 30%. "bmEachStep" and "bmLeadtime" stand for "best member at each step" and "best member over lead time", respectively. See text for further details.

Conclusions

In the running PP-AWARE period, we have applied a lot of verification metrics which are already established (neighbourhood verification) and tested also new verification metrics based on MMI (pseudomember by Johnson et al. (2020)). Especially the latter is quite useful in the SINFONY project. When using a 40 member object ensemble from NWP, nowcasting and combined products, the number of existing objects could become massively huge and not manageable without applying filter methods like pseudomembers.

All above described methods, and some more, are implemented in R-packages predominantly for DWD-internal usage. However, if the packages are well developed, they could be provided to the community. The R-packages are applicable by namelist control but also interactively. We will provide a flexible reading capability. The packages will have a flexible aggregation functionality over different parameters. A visualization via R-Shiny app will give the possibility to interactively visualize and aggregate scores in a way the user desired. Up to now, we do not plan to integrate an extensive pre-processing like regridding or restructuring. We focus only on the computation of the scores and the user has the responsibility to unify the data in advance.

References

1. Davis, C. A., Brown, B. G., Bullock, R., and Halley-Gotway, J.: The Method for Object-Based

Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program, Weather and Forecasting, 24, 1252 – 1267,

1. Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorological Applications, 15, 51–64, https://doi.org/10.1002/met.25, 2008.

2. Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of Spa-tial Forecast Verification Methods, Weather and Forecasting, 24, 1416 – 1430, https://doi.org/ 10.1175/2009WAF2222269.1, 2009.

3. Johnson, A., Wang, X., Wang, Y., Reinhart, A., Clark, A. J., and Jirak, I. L.: Neighborhood- and Object-Based Probabilistic Verification of the OU MAP Ensemble Forecasts during 2017 and 2018 Hazardous Weather Testbeds, Weather and Forecasting, 35, 169 – 191, https://doi.org/10.1175/WAF-D-19-0060.1, 2020.

4. Roberts, N. M. and Lean, H. W.: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, Monthly Weather Review, 136, 78 – 97, https://doi.org/ 10.1175/2007MWR2123.1, 2008.

5. Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., Thomas, K. W., Levit, J. J., Coniglio, M. C., and Wandishin, M. S.: Toward Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Ensemble Membership, Weather and Forecasting, 25, 263 – 280, https://doi.org/10.1175/2009WAF2222267.1, 2010.

6. Skok, G. and Roberts, N.: Estimating the displacement in precipitation forecasts using the Fractions Skill Score, Quarterly Journal of the Royal Meteorological Society, 144, 414–425, https://doi.org/ https://doi.org/10.1002/qj.3212, 2018.

7. Stein, J. and Stoop, F.: Neighborhood-Based Contingency Tables Including Errors Compensation, Monthly Weather Review, 147, 329–344, https://doi.org/10.1175/MWR-D-17-0288.1, 2018.

8. Zeng, Y., Blahak, U., and Jerger, D.: An efficient modular volume-scanning radar forward operator for NWP models: description and coupling to the COSMO model, Quarterly Journal of the Royal Meteorological Society, 142, 3234–3256, 2016.

7 Overview of forecast methods, representation and user-oriented products linked to HIW

Question: How well is HIW represented in postprocessing? What are the pros/cons of DMO vs. PostPro with respect to HIW phenomena predictions? What is the current predictive skill, and the user's interpretation of forecast value in high-impact weather situations (observed and/or forecast)?

HIW phenomena studied: fog/visibility, convection related CW (thunderstorms, lightning, squalls, showers, flash floods)

7.1 Postprocessing vs. direct model output (DMO) for HIW

This task relates to the following key project aspect: How well high-impact weather is represented in postprocessing and modeling. In order to be able to choose the best method to predict HIW, we need to understand the state-of-the art in this field, both in direct modelling and processed model data.

Part 1: Overview of fog forecast

Yu. Khlestova and E. Tatarinovich RHM

The fog is suspended cloud particles in the air near the surface (height 1.5-2 m), which reduces horizontal visibility up to 1 km and less (Khrgian and Mazin, 1989). The main reasons of fog formation are the air mass advection, radiative cooling due to cloudless meteorological conditions, orography effects and anthropogenic activity. Anthropogenic activity stimulates an increase of cloud condensation nuclei number concentration and promotes cloud formation. The most complete fog forecast includes the time of fog formation and duration, its vertical extent and intensity. The fog vertical extent and fog duration depend on atmospheric moisture content and specific meteorological conditions (air stratification, wind speed, cloud amount and structure). Fogs along the vertical extent are divided into ground level (below 2 m), low (2-10 m), medium (1-100 m) and high (more 100 m) fog (Khrgian and Mazin, 1989).

The horizontal visibility (VIS) is the main characteristic of fog intensity. VIS is based on the Koschmieder's formula (Koschmieder H., 1924):

$$VIS = \frac{-\ln(\varepsilon)}{\beta_\lambda}$$

where ε is the eye contrast sensitivity threshold (usually 0.05 or 0.02) (ICAO, 2010; Stoelinga and Warner, 1999), β is the extinction coefficient, λ is the irradiance wavelength, which is usually equal to 550 nm (Trautmann and Bott, 2002). The theoretical formulation of β is based on Mie theory:

$$\beta_\lambda = \int_0^\infty Q_{ext,\lambda} n(r) r^2 \, dr$$

where Q_{ext} is Mie efficiency factor, r is the radius of cloud droplets, n(r) is the number density of cloud droplets (Gultepe and Milbrandt, 2007). The Mie efficiency factor is about 2 for

cloud and rain droplets (Koening, 1971). The theoretical equation of β is not used in the operational forecast. Firstly, the theoretical formulation is expensive for operational weather prediction. Secondly, the theoretical formulation of extinction coefficient requires a more detailed description of cloud droplet's number density. The extinction parameter can also be parametrized using standard meteorological values or microphysical cloud characteristics. There are three main approaches to fog prediction using parametrization. According to the first approach, visibility can be forecasted using empirical relations between β and meteorological parameters (air temperature, dew point temperature, wind speed, air pressure) by observation. Empirical ratios are created for specific points (specific climate and orography) and synoptic situations. This method requires a preliminary analysis of meteorological conditions, since empirical relations are found for specific air conditions and fog physical mechanisms.

The second approach is the use of fog forecasting techniques based on machine learning methods (Abdulkareem K. H. et al., 2019; Zhu et al., 2017; Oguz and Pekin, 2019). The input data is observed or simulated air temperature, dew point temperature, atmospheric pressure, relative humidity, wind speed and direction at 10 m. ML methods organize the forecast based on a set of air condition data. The result is the extinction coefficient or visibility.

According to the third approach, the extinction coefficient can be calculated using β parametrization and numerical weather prediction results (directly in the model or in postprocessing). All parametrizations are obtained based on observations. There are two types of numerical visibility prediction: the meteorological approach and the microphysical approach. The extinction coefficient is based on meteorological characteristics according to the meteorological approach. Examples of "meteorological approach" parametrizations with its applications are shown in Table 18. The T is the air temperature (°C), T_d is the dew point temperature (°C), RH is the relative humidity (%), a_{1-8} are constants based on measurement data. The main limitation of the meteorological approach is that meteorological values are not able to describe the cloud structure, which reduces the forecast accuracy.

Parametrization of β and VIS, km	Source	Application
$\beta = 6000(T - T_d)/RH^{1.75}$	Doran et al., 1999	Forecast System Laboratory
$VIS = a_1 ln(RH) + a_2$ $VIS = a_3 RH^{a_4} + a_5$ $VIS = a_6 RH^2 + a_7 RH + a_8$	Gultepe et al., 2009	
$VIS = 60000exp((-2.5)/80(PH - 15))$	Bang et al., 2009	

Table 18: The meteorological approach of extinction coefficient.

The microphysical approach of β is based on cloud characteristics. The microphysical parametrizations are shown in Table 19. The N_c is the number concentration of cloud droplets (cm^{-3}), N_i is the number concentration of ice particles (cm^{-3}), QC is the liquid water content (g/m^3), QI is the ice water content (g/m^3), R is the radius of cloud droplets (m), b_{1-3} are constants based on measurement data.

The relation (Stoelinga and Warner, 1999) is operatively used for numerical weather forecasting in the WRF model (Weather Research and Forecasting Model). The parametrization of β(Kunkel B.A., 1984) is based on (Eldridge R.G., 1966; Eldridge R.G., 1971; Pinnick et al., 1978; Tomasi and Tampieri, 1976) works and is widely used in HARMONIE (HIRLAM ALADIN Research on Meso-scale Operational NWP In Europe), AROME (Applications of Research to Operations at Mesoscale) and is also applied to the one-dimensional fog forecast

model COBEL. The β-description of the PAFOG fog prediction model is based on (Trautmann and Bott, 2002). The fog forecast in Unified Model uses the method (Clark et al., 2008).

Parametrization of β and VIS, km	Source	Application
$\beta = b_1 QC^{b_2}$	Eldridge R.G., 1966; Eldridge R.G., 1971; Pinnick et al., 1978; Tomasi and Tampieri, 1976	Kunkel B.A., 1984
$\beta = 144.7QC^{0.88}$	Kunkel B.., 1984	COBEL (Muller M.D., 2006); HARMONIE (Kettler T.T., 2020); AROME (Philip et al., 2016); WRF (Creighton et al., 2014); Texeira et al., 2001
$\beta = 163.9QI$	Stoelinga and Warner, 1999	WRF (Creighton et al., 2014); HARMONIE (Kettler T.T., 2020)
$\beta = 230R/QC$	Zverev A.S., 1977	Shatunova et al., 2015
$\beta = 1.5??N_cR^2$	Clark et al., 2008	UM (Claxton et al., 2008; Boutle et al., 2016)
$\beta = b_3 QC^{2/3}(N_c)^{1/3}$	Bott and Trautmann, 2002	PAFOG

Table 19: The microphysical approach of extinction coefficient.

The visibility forecast within the numerical weather prediction can be improved using one-dimensional fog models (1D) and specific settings of model physics. Well-known 1D models are the University of Toulouse COBEL model (Couche Brouillard Eau Liquide) (Bergot and Guedalia, 1994; Muller M.D., 2006; Muller et al., 2007) and the PAFOG (PArameterized FOG) model of the University of Bonn (Bott and Trautmann, 2002; Masbou M., 2008; Mohr et al., 2009). Thermodynamic, radiative and microphysical processes of 1D models are presented with higher vertical resolution, especially in the planetary boundary layer. The lower vertical grid spacing promotes to improve the description of turbulent fluxes and radiative cooling in fog conditions (Trautmann and Bott, 2002a-b).

Thus, the operational VIS prediction is usually based only on one-moment microphysics results (liquid and ice water contents). However, we can also account for the number concentration of particles using two-moment microphysics. The two-moment microphysics implementation and aerosol representation lead to a more sufficient cloud description and fog.

Finally, the detailed tuning of model physical schemes is required to improve the fog forecast. For example, the formation of stable atmospheric stratification is assumed for radiative fog formation, and this is necessary to reduce the errors of the simulated turbulent heat transfer (Thoma and Bott, 2011; Masbou and Bott, 2010). The time of fog formation and dispersion depends on the model description of radiation processes (Antoine S., 2020). Schemes of fog prediction, including the aerosol physical properties and dynamics and cloud-aerosol interaction, show more sufficient results (Vie et al., 2015; Clark et al., 2008).

It can be concluded that the quality of fog prediction depends mainly on the model grid spacing and the approaches of turbulent, microphysical, radiative processes and surface-air exchanges. The fog prediction tasks today have two basic directions. Firstly, we need to decrease the model grid spacing due to the locality and spatial heterogeneity of fog events (Boutle et al., 2016; Philip et al, 2016). And, secondly, the lower grid spacing requires a

revision of model physics, new approaches and description of urban environment (Roebber et al., 2004; Zangl G., 2021).

Taking into account the development of atmospheric modeling, the forecast of the horizontal visibility in the postprocessing (including machine learning methods) seems to be the most appropriate option. It is physically justified, since it is based on the prognostic cloud characteristics and/or parameters of environment. We can apply a set of parametrization (ensemble) of horizontal visibility (Tables 18 and 19). The fog is a meteorological phenomenon with a high degree of locality. We try to reduce the probability of prognostic error by using a set of postprocessing approaches. The main issue of the development of these methods is the lack of instrumental measurements of visibility due to the specifics of observations.

References

1. International Civil Aviation Organization. (2010). Meteorological Service for International Air Navigation. Annex 3 to the Convention on International Civil Aviation. Part I Core SARPs. Part II Appendices and Attachments. ICAO.

2. Creighton, G., Kuchera, E., Adams-Selin, R., McCormick, J., Rentschler, S., & Wickard, B. (2014). AFWA diagnostics in WRF.

3. Kettler, T. T. (2020). Fog forecasting in HARMONIE: a case study to current issues with the overestimation of fog in HARMONIE. (Master's thesis).

4. Khrgian, A. Kh., & Mazin, I. P. (Eds). (1989). Clouds and cloudy atmosphere. The handbook. Hydrometeoizdat, Leningrad. In Russian.

5. Abdulkareem, K. H., Mohammed, M. A., Gunasekaran, S. S., Al-Mhiqani, M. N., Mutlag, A. A., Mostafa, S. A., ... & Ibrahim, D. A. (2019). A review of fog computing and machine learning: concepts, applications, challenges, and open issues. IEEE Access, 7, 153123-153140.

6. Bang, C. H., Lee, J. W., & Hong, S. Y. (2008). Predictability experiments of fog and visibility in local airports over Korea using the WRF model. Journal of Korean society for atmospheric environment, 24(E2), 92-101.

7. Bergot, T., & Guedalia, D. (1994). Numerical forecasting of radiation fog. Part I: Numerical model and sensitivity tests. Monthly Weather Review, 122(6), 1218-1230.

8. Bott, A., & Trautmann, T. (2002). PAFOG—A new efficient forecast model of radiation fog and low-level stratiform clouds. Atmospheric Research, 64(1-4), 191-203.

9. Boutle, I. A., Finnenkoetter, A., Lock, A. P., & Wells, H. (2016). The London Model: forecasting fog at 333 m resolution. Quarterly Journal of the Royal Meteorological Society, 142(694), 360-371.

10. Clark, P. A., Harcourt, S. A., Macpherson, B., Mathison, C. T., Cusack, S., & Naylor, M. (2008). Prediction of visibility and aerosol within the operational Met Office Unified Model. I: Model formulation and variational assimilation. Quarterly Journal of the Royal Meteorological Society, 134(636), 1801-1816.

11. Claxton, B. M. (2008). Using a neural network to benchmark a diagnostic parametrization: the Met Office's visibility scheme. Quarterly Journal of the Royal Meteorological Society, 134(635), 1527-1537.

12. Doran J.A., Roohr, P.J., Beberwyk, D.J., Brooks, G.R., Gayno, G.A., Williams, R.T., Lewis, J.M. & Lefevre, R.J. (1999). The MM5 at the air Force Weather Agency – New products to support military operations. The 8^{th} Conference on Aviation, Range, and Aerospace Meteorology, Dallas, Texas, 10-15 January.

13. Eldridge, R.G., (1966). Haze and fog aerosol distributions. J. Atmos. Sci., 28, 605-613.

14. Eldridge, R.G., (1971). The relationship between visibility and liquid water content in fog. J. Atmos. Sci28, 1883-1186.

15. Gultepe, I., & Milbrandt, J. A. (2007). Microphysical observations and mesoscale model simulation of a warm fog case during FRAM project. In Fog and Boundary Layer Clouds: Fog Visibility and Forecasting (pp. 1161-1178). Birkhäuser Basel.

16. Gultepe, I., Pearson, G., Milbrandt, J. A., Hansen, B., Platnick, S., Taylor, P., ... & Cober, S. G. (2009). The fog remote sensing and modeling field project. Bulletin of the American Meteorological Society, 90(3), 341-360.

17. Kahraman, O. Ğ. U. Z., & Pekin, M. A. (2019). Predictability of fog visibility with artificial neural network for Esenboga airport. Avrupa Bilim ve Teknoloji Dergisi, (15), 542-551.

18. Koening, L.R. (1971). Numerical experiments pertaining to warm-fog clearing. Mon. Wea. Rev., 9, 227-241.

19. Koschmieder, H. (1924). Theorie der horizontalen Sichtweite. Beitrage zur Physik der freien Atmosphare, 33-53. In German.

20. Kunkel, B. A. (1984). Parameterization of droplet terminal velocity and extinction coefficient in fog models. Journal of Applied Meteorology and Climatology, 23(1), 34-41.

21. Masbou, M. (2008). LM-PAFOG: a new three-dimentional fog forecast model with parametrised microphysics (Doctoral dissertation, Université Blaise Pascal-Clermont-Ferrand II). 188 pp.

22. Mohr, C., Alberts, I., Masbou, M., & Bott, A. (2009). Nebelbildung am Flughafen München: Klimatologie und Modellierung. Report Universität Bonn. 77 pp.

23. Müller, M. D., Masbou, M., Bott, A., & Janjic, Z. (2005, September). Fog prediction in a 3d model with parametrized microphysics. In WWRP Int. Symp. on Nowcasting and Very Short-Range Forecasting, WWRP.

24. Müller, M.D., Schmutz, C., & Parlow, E. (2007). A One-Dimensional Ensemble Forecast and Assimilation System for Fog Prediction. Pure and Apllied Geophysics, 164, 1241-1264.

25. Philip, A., Bergot, T., Bouteloup, Y., & Bouyssel, F. (2016). The impact of vertical resolution on fog forecasting in the kilometric-scale model arome: a case study and statistics. Weather and Forecasting, 31(5), 1655-1671.

26. Pinnick, R.G., Hoihjelle, D.L., Fernandez, G., Stenmark E.B., Lindberg J.D., Hoidale G.B., & Jenings, S.G. (1978). Vertical structure in atmospheric fog and haze and its effect on visible nd infrared extinction. J. Atmos. Sci., 35, 2020-2032.

27. Roebber, P.J., Schultz, D.M., Colle, B.A., & Stensrud, D.J. (2004). Toward improved prediction: High-resolution and ensemble modeling systems in operations. Weather and forecasting, 19(5), 936-949.

28. Shatunova, M. V., Rivin, G. S. & Rozinkina, I. A., (2015). Visibility forecasting for February 16-18, 2014 for the region of the Sochi-2014 Olympic Games using the high–resolution COSMO-Ru1 model. Russian Meteorology and Hydrology, 40(8), 523-530.

29. Stoelinga, M. T., & Warner, T. T. (1999). Nonhydrostatic, mesobeta-scale model simulations of cloud ceiling and visibility for an East Coast winter precipitation event. Journal of Applied Meteorology and Climatology, 38(4), 385-404.

30. Teixeira, J., & Miranda, P. M. (2001). Fog prediction at Lisbon airport using a one-dimensional boundary layer model. Meteorological Applications, 8(4), 497-505.

31. Tomasi, C., & F. Tampieri, (1976). Features of the proportionality coefficient in the relationship between visibility and liquid water content in haze and fog. Atmosphere, 14, 61-76.

32. Trautmann, T., & Bott, A. (2002a). A numerical model for local fog prediction. Part 1: Parameterized microphysics and radiation. Scientific reports from the Institute of Meteorology at the University of Leipzig, 26, 1-15. In German.

33. Trautmann, T., & Bott, A. (2002b). A numerical model for local fog prediction. Part 2: Handling of soil and vegetation. Scientific reports from the Institute of Meteorology at the University of Leipzig, 26, 16-30. In German.

34. Zangl, G. (2021). News on ICON-MWP. Recent model improvements and systematic resolution-dependence tests to determine needs for further development. 23^{rd} COSMO General Meeting, 14-17 September (teleconference).

35. Zhu, L., Zhu, G., Han, L., & Wang, N. (2017). The application of deep learning in airport visibility forecast. Atmospheric and Climate Sciences, 7(03), 314.

36. Zverev A.S., (1977). Synoptic meteorology. Hydrometeoizdat, Leningrad. In Russian.

Part 2: Tornado hazard prediction with COSMO-Ru parameters and indices

Denis Zakharchenko, RHM

Recent research [Chernokulsky et.al, 2020] showed that on average Russia experiences from 100 to 150 tornadoes per year, although during some years this number can rise up to 350 events. Although the majority of these tornadoes are considered non-significant, about 10 percent of twisters can reach F-2 [Fujita, 1971] (or EF-2 [McDonald, et.al, 2004]) rated intensity and higher, causing serious damage and human deaths and injuries. With very few exceptions, these significant tornadoes are associated with deep persistently rotating updrafts, found within supercells [Doswell & Burgess, 1993] and mesoscale convective systems.

Current numerical weather prediction models accepted by RosHydroMet cannot resolve tornadic vortices in operative forecasts. Although, a number of indices and parameters based on simulated characteristics of deep convection systems can help to predict the risk of a significant tornado forming along with accompanying severe weather hazards.

The current operative configuration for the COSMO-Ru model allows to run simulations with 2.2 km horizontal grid spacing within a domain covering the European Part of Russia (EPR) and other eastern European countries (Figure 94). Simulations with this 2.2-km grid configuration are performed with the Latent Heat Nudging (LHN) application. Parametrization schemes for COSMO-Ru domains are listed in Table 18.

The model output data includes the parameters for severe weather hazard diagnosis and prediction, such as the Lightning Potential Index (LPI) [Yair et.al, 2010], Hailcast Parameters [Adams-Selin & Ziegler, 2016], Supercell Detection Index (SDI) [Wicker et.al, 2005] along with common convective instability indices (CAPE, CIN).

The Supercell Detection Index represents a useful tool for identifying rotating updrafts in simulated convective cells and systems. Though the formation of a significant tornado requires the existence of mesocyclonic updrafts, it is not a sufficient condition, and other tornadogenesis factors must be taken into account. It is noted by the authors, that a value of 0.003 1/s is considered a significant threshold for supercell storms. It is important to note, that the SDI has two variations. In this study we refer to SDI_2. In this index, the positive values represent counterclockwise rotation (meso-cyclonic) whilst the negative represent meso-anticyclonic clockwise rotation.

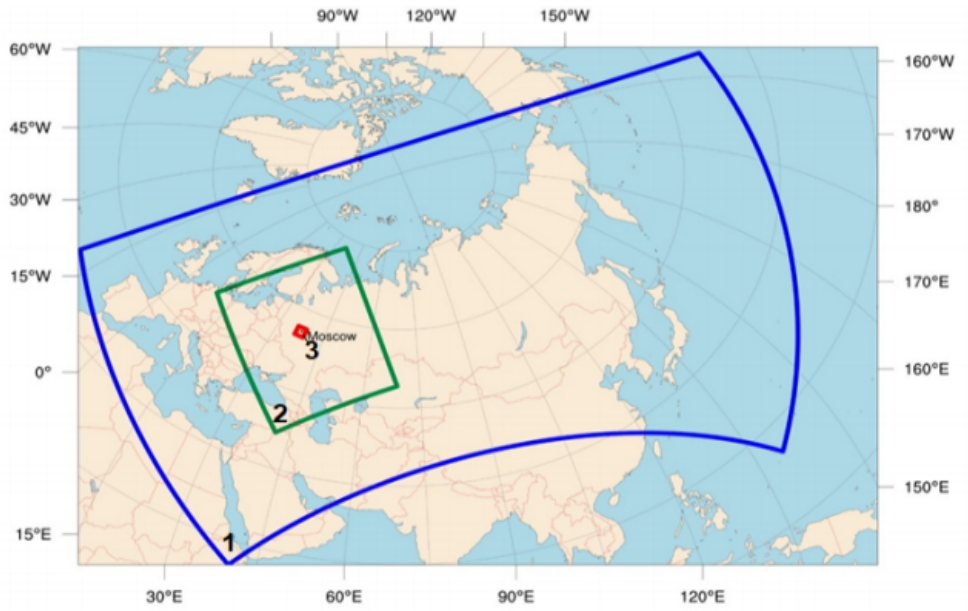


Figure 94: Operative domains of the COSMO-Ru NWP system: 1- 6.6 km grid, 2- 2.2-km grid, 3- 1.1-km grid.

Grid spacing	**6.6 km**	**2.2 km**	**1.1 km**
Vertical layers	40	50	
Mircophysics scheme	Two-category ice scheme	Three-category ice scheme	
Convection scheme	Mass flux Tiedke scheme	Mass flux Tiedke scheme (Shallow convection scheme)	
Turbulence scheme	1-D TKE based diagnostic closure		
Time step [s]	50	20	5

Table 20: Preferences for COSMO-Ru operative domains.

Another parameter for estimating conditions favorable for significant tornado formation is the Significant Tornado Parameter (STP) [Thompson et.al, 2003]. It has not been implemented in the model, but can be calculated using the simulated instability indices and wind fields. This parameter requires the 0-1 km layer Storm-Relative helicity, which is calculated for right-mover (meso-cyclonic) tornadic supercells using the Bunkers storm mo-

tion method [Bunkers et.al, 2000]. It is specified that STP values higher than 1 represent a hazard for significant tornado occurrence. However, the disadvantage of this parameter as noted by the authors is in a high false alarm rate.

In previous studies with COSMO-Ru it has been noted, that the STP marks vast areas of favorable conditions for significant tornado formation, though often not within the range of simulated convective cells.

In respect to the statements above, the Significant Tornado Parameter (STP) and the Supercell Detection Index (SDI) simulated fields can be more informative if examined in complex.

Figure 95: Occurrence of Tornadoes in the European part of Russia in 2021.

The idea of this study is to analyzed simulated COSMO-Ru fields during severe weather outbreaks with tornado activity in the European part of Russia in 2021. According to the European Severe Weather Database [ESWD], there were 85 tornado records in Russia during the year, with 45 events identified as waterspouts, mostly observed on the coasts of the Black sea (Figure 95).

The most significant events took place on May 15 and August 2 the central European part of Russia. Our case studies are based on these two events.

On May 15, 2021 a group of supercells and mesoscale convective systems travelled across Moscow, Vladimir, Yaroslavl, Ivanovo, Kostroma and Vologda regions, resulting in a tornado outbreak and widespread straight-line wind damage. Analysis of satellite imagery of forest damage revealed at least 6 tornado tracks and a 360-km long squall damage path, indicating a possible derecho event in Vladimir and Kostroma regions (Figure 96). Several tornadoes, including one originated in a supercell in Yaroslavl region are rated F-2 and considered significant.

The initial time for the COSMO-Ru 2.2km simulation is 0:00 UTC May 15, 2021. At 08:00 UTC simulated convective cells already develop along the frontline, following a NNE trajectory and become recognizable in the simulated radar reflectivity field. In several cells the SDI already reveals signs of rotation, including a meso-cyclonic and meso-anticyclonic

couple of updrafts in Yaroslavl region, which can be an indicator of a process known as "Supercell splitting". At the moment, a stretched and narrow area of STP values exceeding the thresholds is present to the east of the simulated supercells with a maximum of 6 located in Ivanovo region (Figure 97).

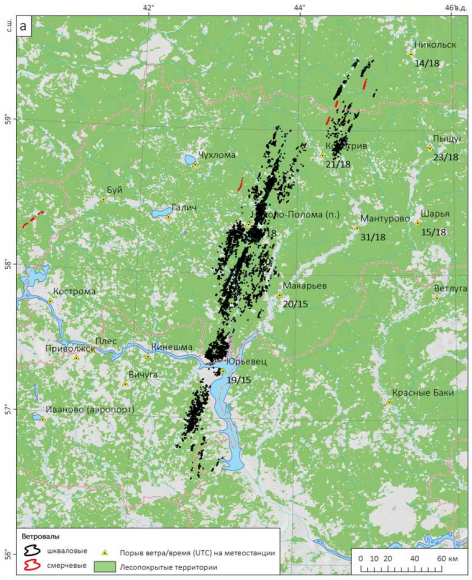


Figure 96: Black hatching represents squall forest damage, red hatching – tornado-induced damage.

At 13:00 UTC the area of increased STP values breaks into separate fragments, one of which migrates to Yaroslavl region. Approximately at this time an EF-2 tornado hit the town of Lyubim in the northeastern part of the region. In the simulation, a cell with high SDI values is located just at the northern border of the region, though not crossing the area with the highest STP values at the time (Figure 5).

Between 15:00 and 16:00 UTC a simulated mesoscale convective system with several rotating updrafts passes over the Kostroma region, matching the time the derecho event and tornadoes were observed. The STP field showed a maximum of 8 approximately at the location where tornado-induced forest damage were found. The maximum wind gust potential field revealed a vast area of high gust speed potential values, exceeding 38 m/s in some places (Figure 99).

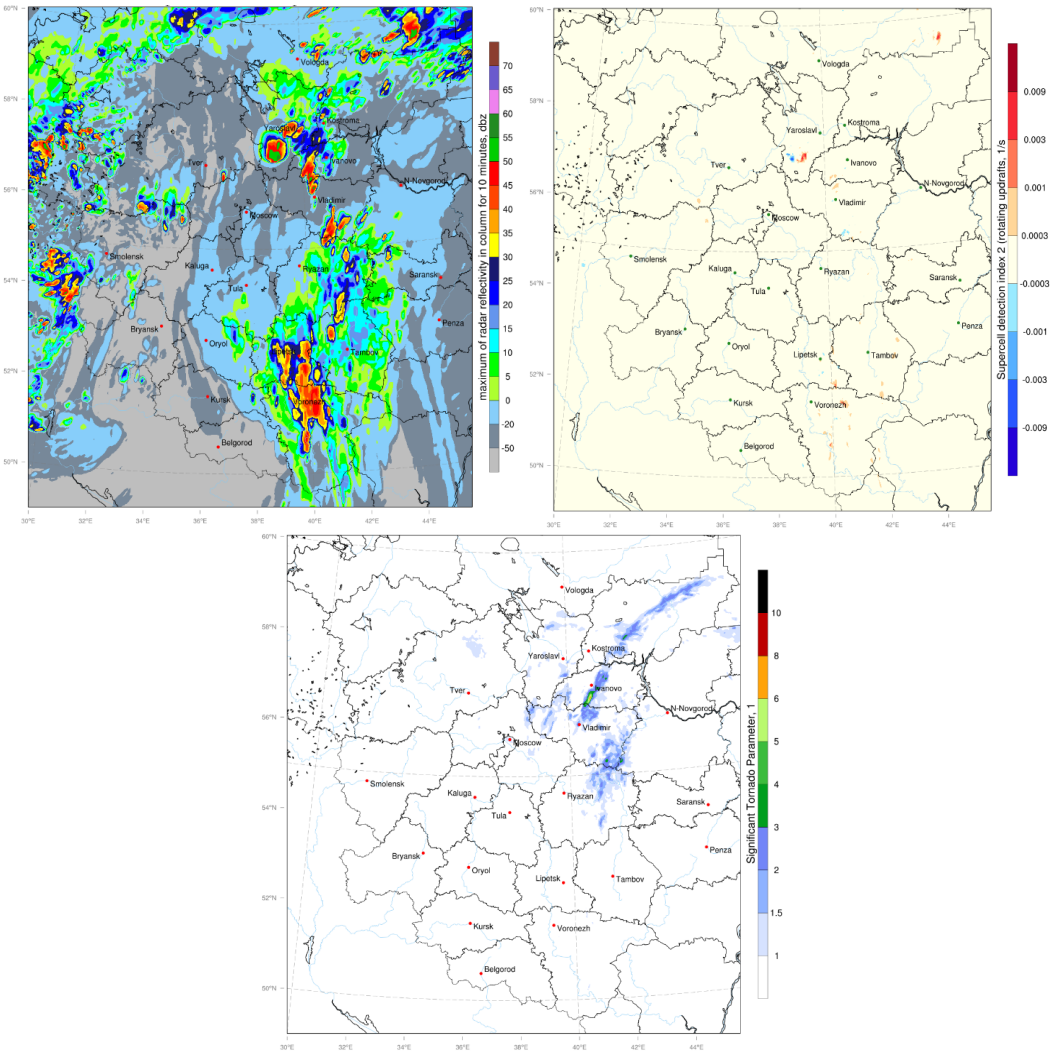


Figure 97: Simulated 2.2-km grid COSMO-Ru fields at 08:00 UTC May 15, 2021: Top left: Radar Reflectivity (dBz). Top right: Supercell Detection Index 2 (SDI_2). Bottom: Significant Tornado Parameter (STP).

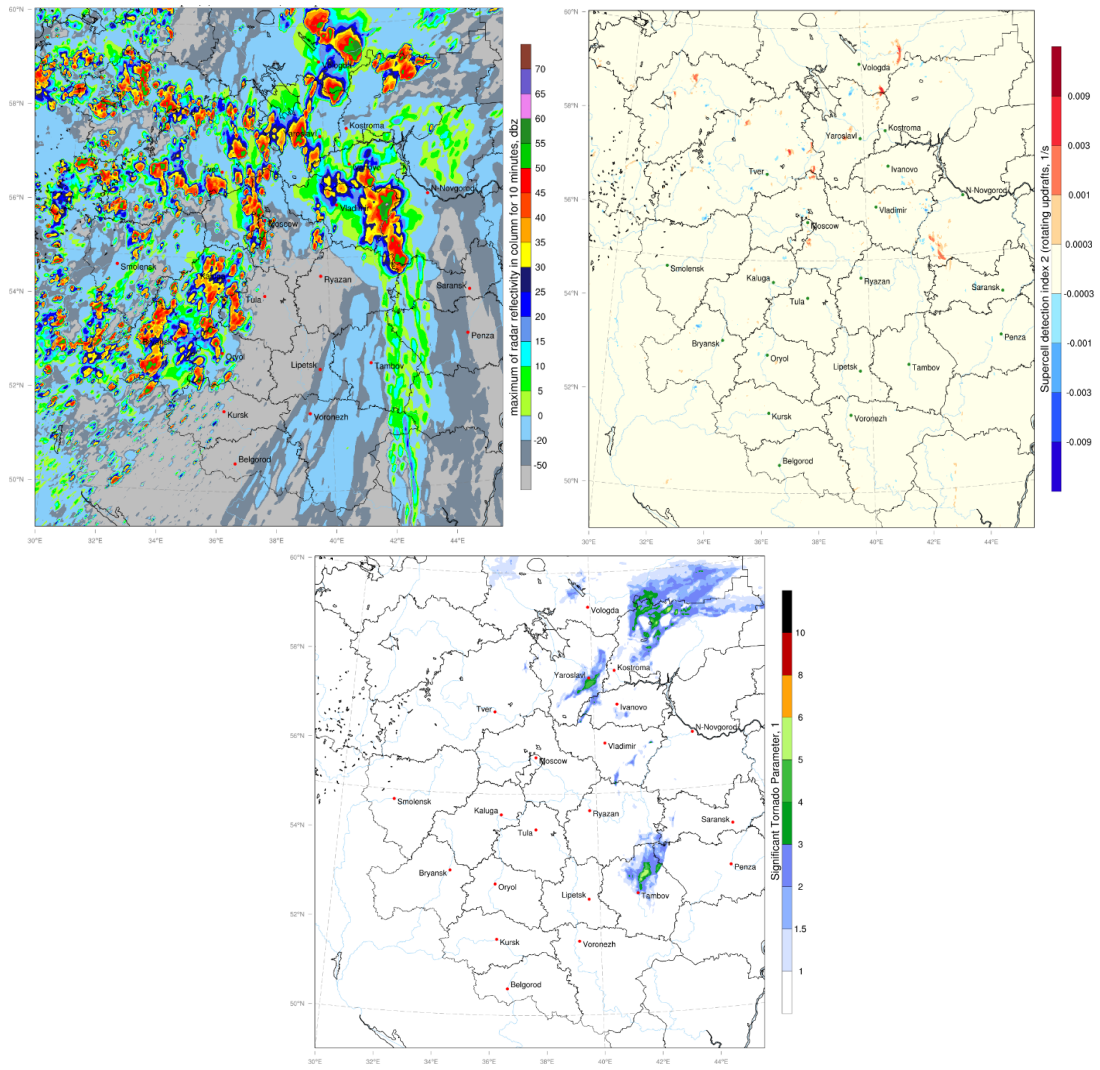


Figure 98: Simulated 2.2-km grid COSMO-Ru fields at 13:00 UTC May 15, 2021: Top left: Radar Reflectivity (dBz). Top right: Supercell Detection Index 2 (SDI_2). Bottom: Significant Tornado Parameter (STP).

In order to examine the structure of the simulated mesoscale convective system in detail, the 2.2-km grid model data was used to initialize a 1.1-km grid simulation within a small domain, surrounding the path of the derecho. As a result, the simulated radar reflectivity field revealed a “nearly textbook” picture of a bow echo system evolution with a detailed pattern-following area of severe wind gusts exceeding 40 m/s with a distinct gust front (Figure 100).

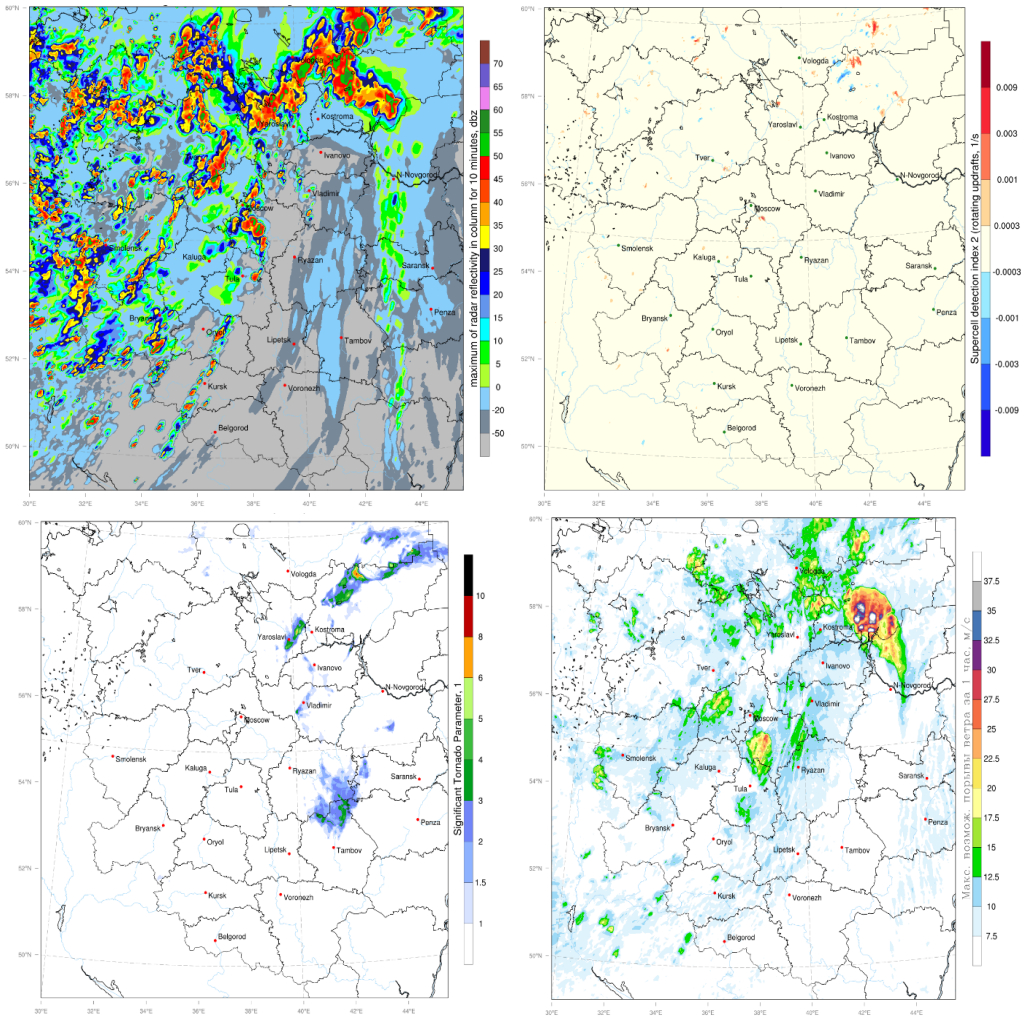


Figure 99: Simulated 2.2-km grid COSMO-Ru fields at 15:00 UTC May 15, 2021: Top left: Radar Reflectivity (dBz). Top right: Supercell Detection Index 2 (SDI_2). Bottom: Significant Tornado Parameter (STP). Bottom right: Maximum 10m AGL wind gust (m/s) for the last forecast hour.

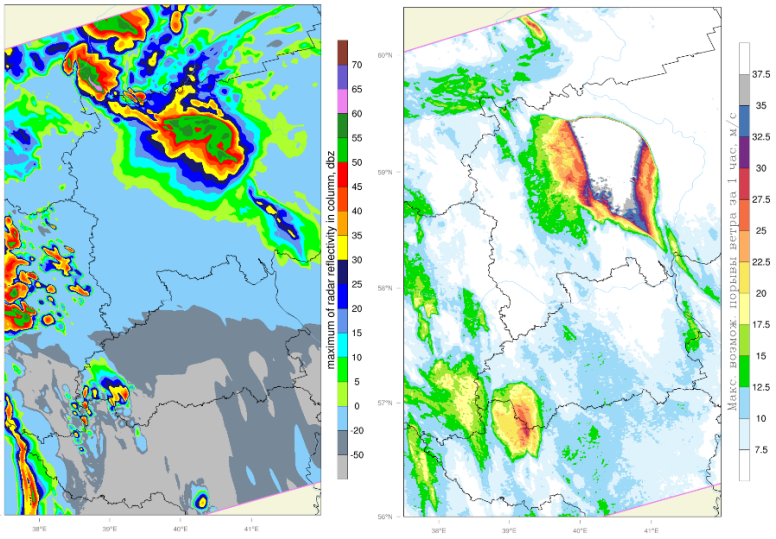


Figure 100: Simulated 1.1-km grid COSMO-Ru fields at 15:00 UTC May 15,2021. Left: Radar reflectivity. Right: Maximum 10m AGL wind gust (m/s) for the last forecast hour.

The other case study event is the August 2 tornado outbreak in Tver and Pskov regions in western Russia. Satellite imagery analysis showed that at least 15 tornadoes touched down during the outbreak, including an F-3 rated tornado hitting Andreapol town in Tver region, killing 3 people and injuring 10 more [ESWD].

Figure 101: Left: The Andreapol F-3 Tornado before hitting the town; Right – Tornado-induced damage in Andreapol town [https://vk.com/meteodnevnik].

As in the previous case study, the model is initialized at 0:00 UTC. The 2.2-km grid simulations revealed the growth of a family of convective cells at 10:00 UTC over the western part of Russia. At 12:00 UTC numerous cells in Pskov and Tver region show signs of rotation and are marked with high SDI values. At the time, a widespread area of increased STP values is present over the surrounding regions with maximum values above 8 in Pskov region (Figure 102). According to ESWD, the tornadoes in Pskov region occurred between 12:00 and 13:00 UTC and reached F-1 intensity.

At 14:00 UTC the STP maximum values, accompanied by several simulated convective cells with high SDI values are observed in Tver region, matching the area where the number of tornadoes, including the Andreapol event occurred between 14:00 and 16:00 UTC (Figure 103).

As in case study 1, a 1.1-km resolution simulation within a smaller domain was initialized. As a result – more distinct radar reflectivity supercell shapes with higher values of SDI

were acquired (Figures 104-105).

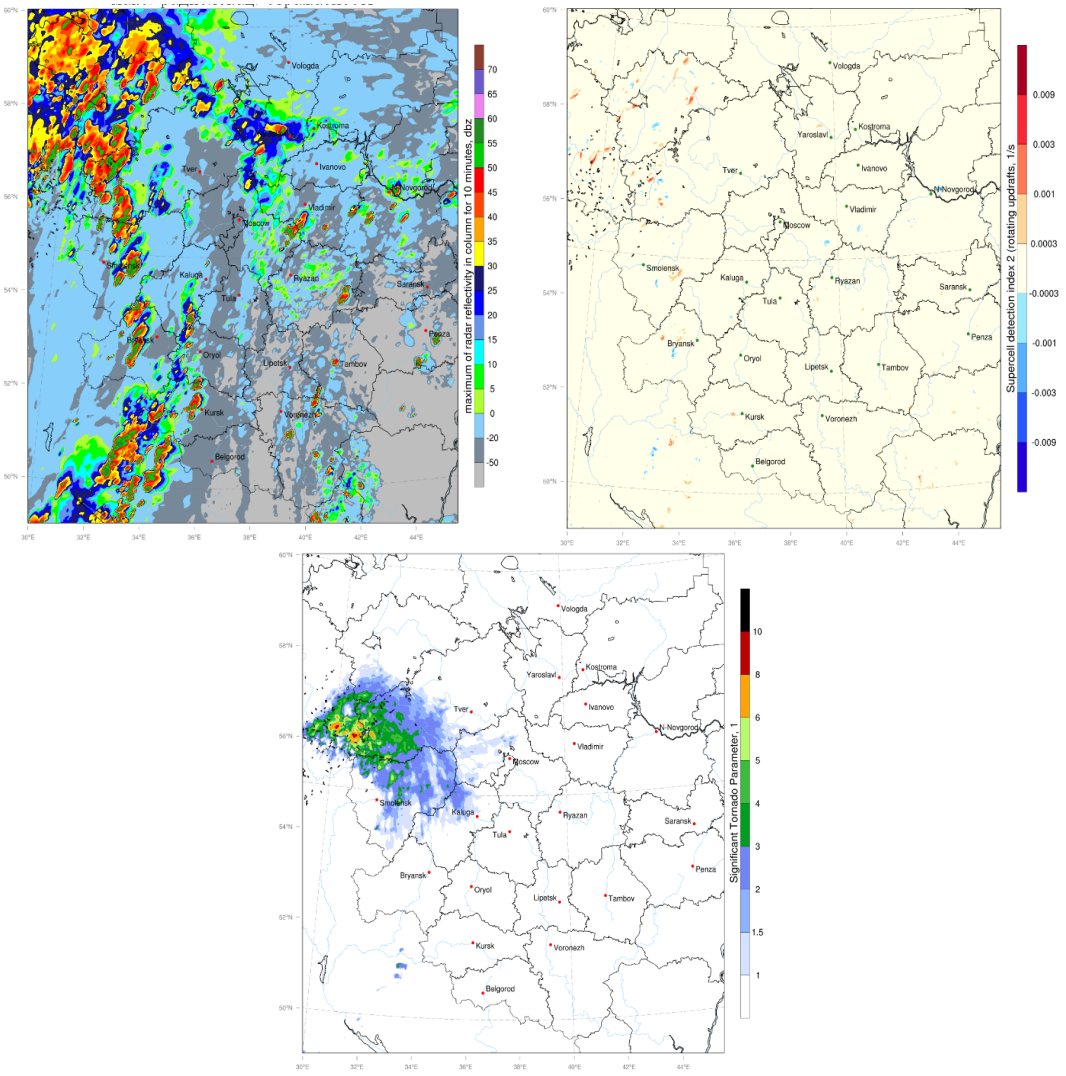


Figure 102: Simulated 2.2-km grid COSMO-Ru fields at 12:00 UTC August 2, 2021: Top left: Radar Reflectivity (dBz).Top right: Supercell Detection Index 2 (SDI_2). Bottom: Significant Tornado Parameter (STP).

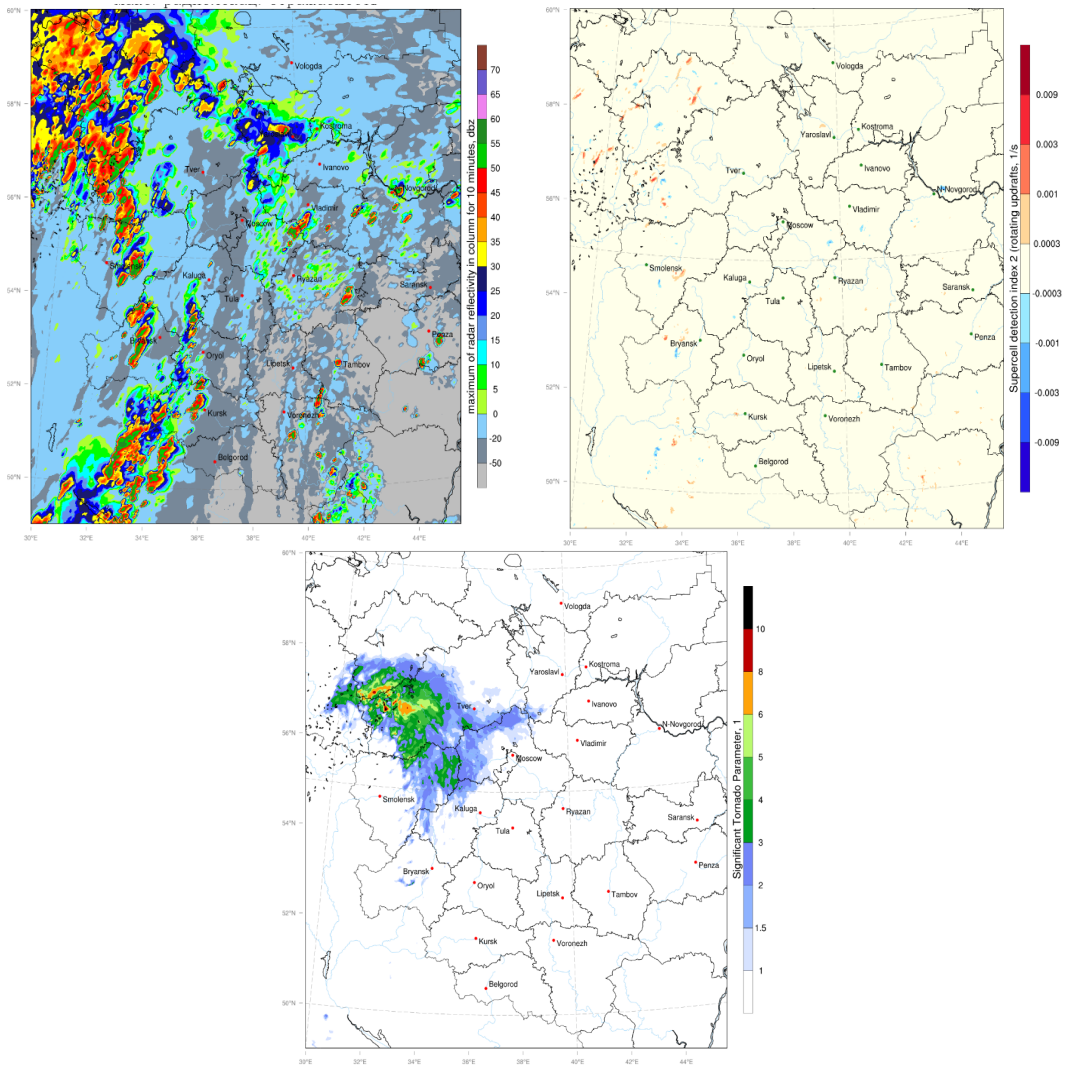


Figure 103: Simulated 2.2-km grid COSMO-Ru fields at 14:00 UTC August 2, 2021: Top left: Radar Reflectivity (dBz).Top right: Supercell Detection Index 2 (SDI_2). Bottom: Significant Tornado Parameter (STP).

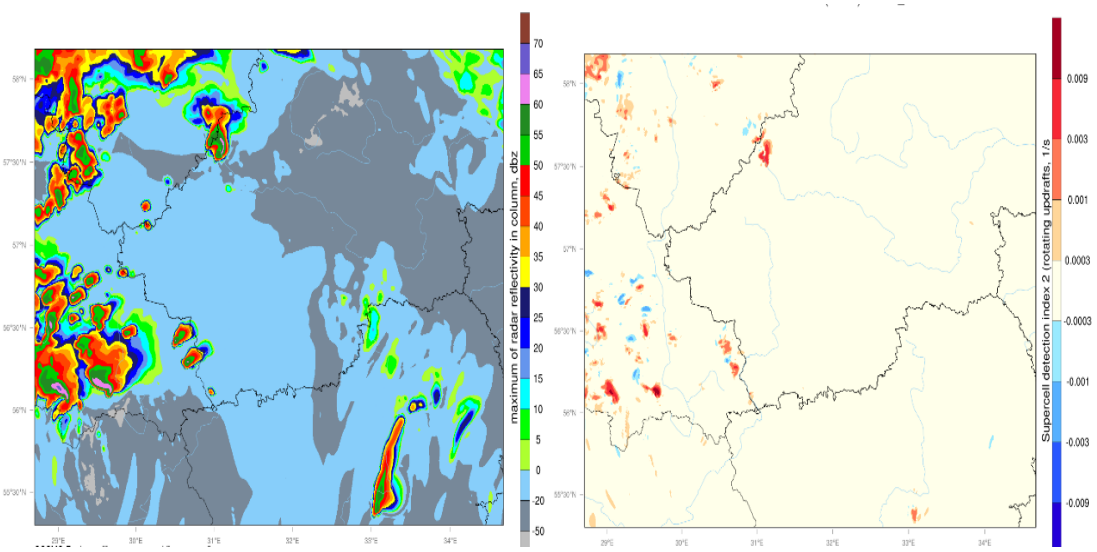


Figure 104: Simulated 1.1-km grid COSMO-Ru fields at 12:00 UTC August 2, 2021: Left: Radar Reflectivity; Right – Supercell Detection Index 2 (SDI_2).

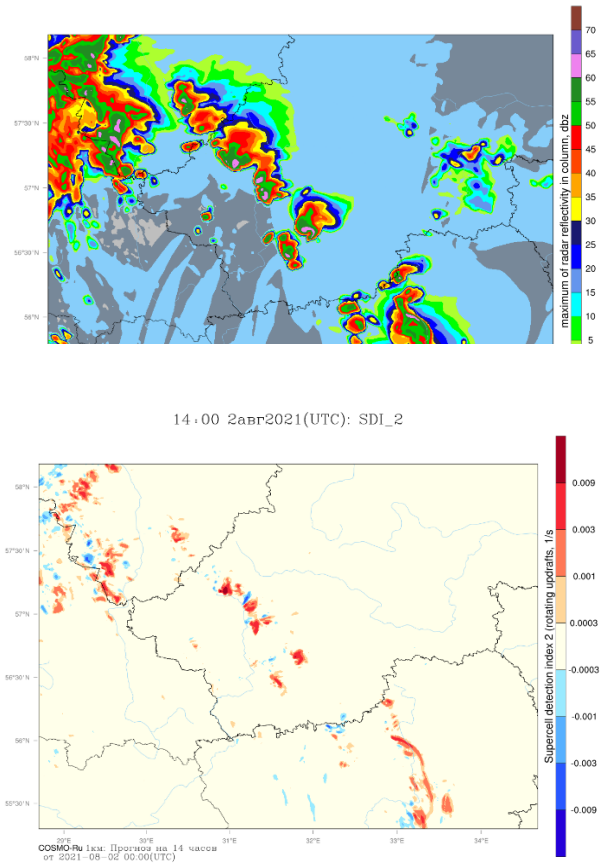


Figure 105: Simulated 2.2-km grid COSMO-Ru fields at 14:00 UTC August 2, 2021: Top left: Radar Reflectivity (dBz).Top right: Supercell Detection Index 2 (SDI_2). Bottom: Significant Tornado Parameter (STP).

In conclusion: predicting tornado occurrence remains a challenge, considering today's op-

erational numerical weather prediction systems and the complex and unexplored nature of these intense phenomena. Nevertheless, convection-permitting models and grids can resolve both dynamical and empirical indicators of certain significant tornado events.

The 2021 case study analysis showed that despite definite spatial and temporal errors when predicting actual significant tornado hazard locations, the joint use of COSMO-based SDI and STP indices can significantly clarify the risk area and exclude an amount of false alarms in certain regions.

It is worth noting, that the 1.1-km-resolution simulations revealed, beyond any doubt, a more detailed picture of the simulated convective cells and systems. This detail may appear crucial for tornadic event diagnosis. Hence, a possibility of an operative 1-km grid COSMO-Ru setup for severe weather prediction is in need of consideration.

References:

1. Adams-Selin, R.D. and C.L. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF.Mon. Wea. Rev.,144, 4919–4939.
2. Bunkers, M. J., Klimowski, B. A., Zeitler, J. W., Thompson R. L., Weisman, M. L., 2000: Predicting Supercell Motion Using a New Hodograph Technique. Weather and Forecasting, 15.
3. Chernokulsky, A., Kurgansky, M, Mokhov, I., Shikhov, A., Azhigov, I., Selezneva, E., Zakharchenko, D., Antonescu, B., Kuhne, T., 2020: Tornadoes in Northern Eurasia: From the Middle Age to the Information Era, Monthly Weather Reviews 148 (8): 3081-3110.
4. Doswell, C., Burgess, D., 1993: Tornadoes and Tornadic Storms: a Review of Conceptual Models. The Tornado: Its Structure, Dynamics, Prediction, and Hazards, Volume 79.
5. Fujita, T., 1971: Proposed characterization of tornadoes and hurricanes by area and intensity. University of Chicago SMRP Research Paper 91, 42.
6. McDonald, J. R., Mehta, K., C., 2004: A Recommendation for an Enhanced Fujita Scale. Texas Tech University, Lubbock, TX, USA.
7. Thompson, R. L., Edwards, R., Hart, J. A. Close proximity soundings within supercell Environments Obtained from the Rapid Update Cycle, 2003 // Weather and Forecasting, 2003, Vol. 18, P. 1243-1261.
8. Wicker, L., Kain, J., Weiss, S., and Bright, D. A Brief Description of the Supercell Detection Index (2005)
9. Y.Yair, B. Lynn, C. Price et al. Predicting the potential for lightning activity in Mediterranean storms based on the weather research and forecasting (WRF) model dynamic and microphysical fields, Journal of Geophysical Research, 2010, Vol. 115, article D04205

7.2 Improving existing post-processing methods

Andrzej Mazur, Grzegorz Duniec

Institute of Meteorology and Water Management – National Research Institute

This work was continued within PP MILEPOST

Introduction

In contrary to other sub-tasks (e.g. 2.1, 3.1) in the Priority Project, the main goal in this activities was the verification against observations of various post-processed results. It means that the effectiveness of post-processing is assessed, not the FR parametrization itself as follows:

$$FR = \left(\frac{W}{14.66}\right)^{4.54}$$

with W being updraft velocity, calculated as

$$W = 0.3 \cdot \sqrt{2 \cdot CAPE}$$

As it was already used in sub-tasks 2.1 and 3.1, FR is to be limited with the temperatures of top/bottom cloud temperatures, CTT and CBT, respectively

$$if\ CTT > -15°C\ FR = FR \cdot \left[max\left(0.01, \frac{-CTT}{15.0}\right)\right]$$

and

$$if\ CBT < -5°C\ FR = FR \cdot \left[max\left(0.01, \frac{15.0 + CBT}{10.0}\right)\right]$$

And again, another limitation is due to lack of convective clouds – if (forecasted) cloud cover is below 25%, FR is set equal to zero. Moreover, case was selected to verification if (for both observations and forecasts) maximum value over the entire domain was greater than 20 strikes/hour , and the duration of the storm was greater than 6 hours.

Observation data (intercloud- and cloud-to-ground lightnings) came from the Polish lightning detection network PERUN, covering Poland and nearest vicinity - parts of neighbouring countries.

The quality of (any) post-processing used in the study was assessed via continuous verification - MAE, RMSE - only. Methods using contingency table nor other discrete verification methods were not used.

Methods

Various methods of post-processing used in the study essentially belonged to the class of Least Mean Squares (LMS) methods and the Artificial Neural Network (ANN) method.

1. *Multi-Linear Regression (MLR)* – class of LMS method with multidimensional input data vector, yet constant over time. Marking corrected forecasts as y, DMO (Direct Model Output) as h, and weight values (to be determined) as b, the method diagram looks as follows.

$$\hat{y}_n = \sum_{j=0}^{J} b_j h_{nj}$$

Solution for regression coeffs

$$\bar{b} = (\hat{H}^T \hat{H})^{-1} \hat{H}^T \bar{y},$$

where

$$\hat{H} = \begin{bmatrix} 1 & h_{11} & \dots & h_{1J} \\ 1 & h_{21} & \dots & h_{2J} \\ \dots & \dots & h_{nj} & \dots \\ 1 & h_{N1} & \dots & h_{NJ} \end{bmatrix}; \bar{y} = \begin{bmatrix} y_0 \\ y_1 \\ \dots \\ y_N \end{bmatrix}; \bar{b} = \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_J \end{bmatrix}$$

Figure 106: Flow chart of the MLR procedure.

1. *Adaptive/Recursive LMS methods.* Basic scheme of the method is presented below. The most important here – from the post-processing point of view – was the forgetting factor λ, that described how long older data should be "remembered".

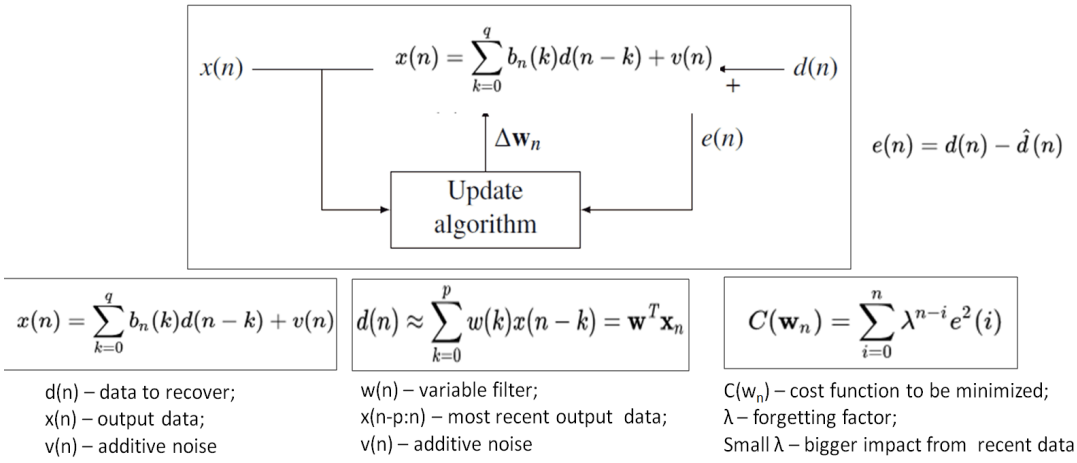


Figure 107: Flow chart of the RLMS procedure.

1. Artificial Neural Networking (ANN) – dealing with post-processing for both EPS- and deterministic forecasts

Inputs to the net were, apart from (time lagged) values of DMO, geographical coordinates λ,φ, and t_s,t_c – lead time and current time of forecast. Basic idea of the ANN is presented in the diagram below. Transfer function was assumed linear, activation function – hyperbolic tangent. The main factor that was modified in the assessment process was the number of hidden neutrons of the net.

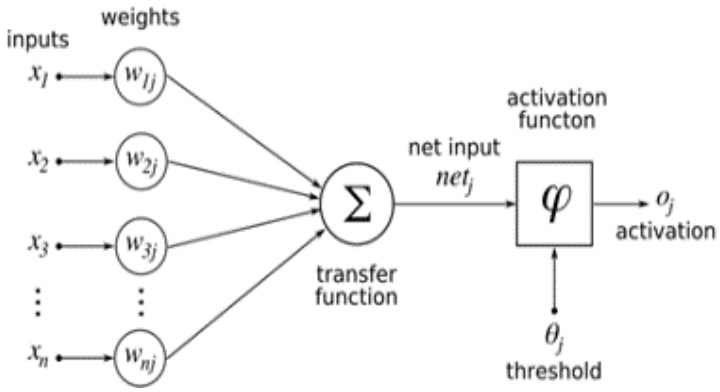


Figure 108: Schematic depiction of ANN.

Space lag (cross-) correlation approach

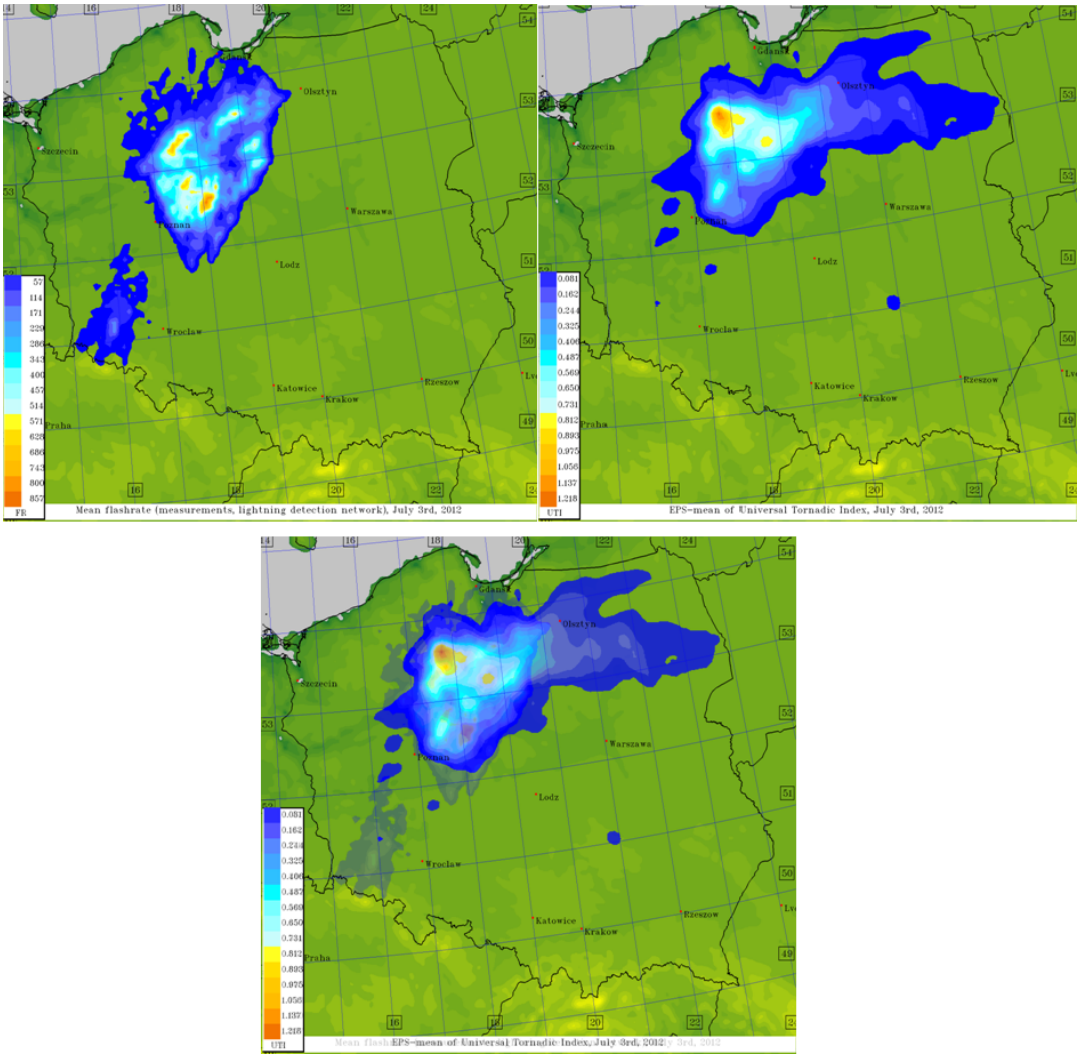


Figure 109: Introduction to cross-correlation procedure.

Similarly, as in subtask 2.1/3.1 cross-correlation procedure was applied. To remind a basic

idea of the approach: when overlap the upper left (observations field) and the upper right (forecasts) panels (Fig. 109), in most cases they do not match (lower panel, Fig. 109). It is possible to improve the forecast by using the cross-correlation (or space lag correlation) method. To do this (using the example from the figure above) one should:

1. Calculate coordinates of "centres of mass" for both distribution patterns (observations vs. forecasts).
2. Compute vector of displacement (VOD) of forecasts to observations as a difference of the two above.
3. Displace linearly every value of forecasts field by the vector of displacement.

In operational work, VOD is calculated from previous model runs (as compared to observations). It is then assumed to remain constant throughout the next run.

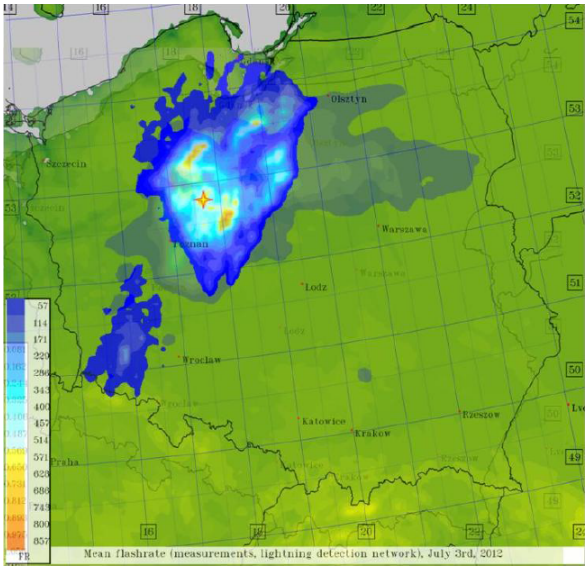


Figure 110: Result of cross-correlation procedure.

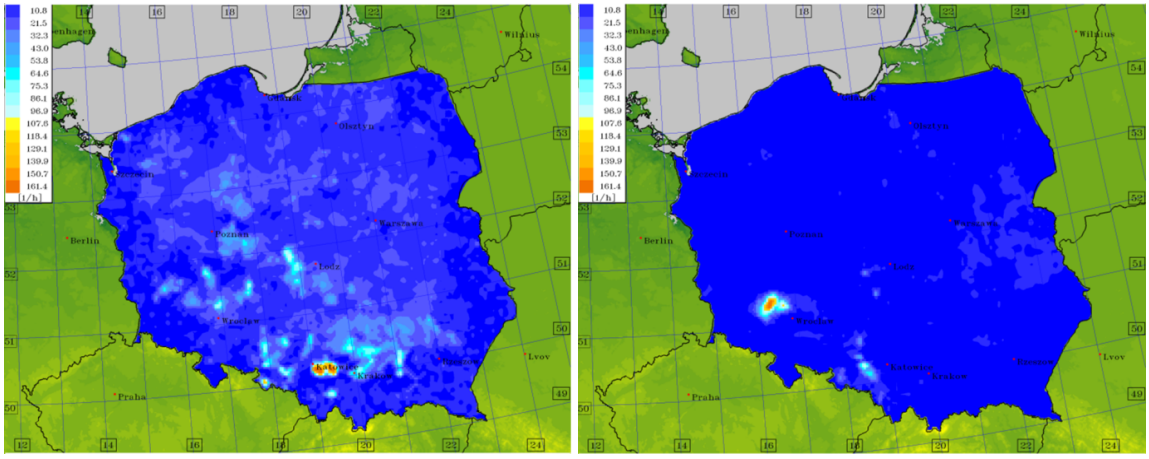


Figure 111: Sample values of (observations – forecasts) for flash rate (lightning frequency). Left - direct model output results, right panel - corrected with VOD procedure.

Examples and detailed results

Various set-ups of post-processing of various methods have been tested over the seven-years period (2011-2017). The learning/testing period: 2011-2016 and the entire 2017 as period for verification were selected. The following table lists the Mean Error, Mean Absolute Error and Root Mean Square Error values for the various set-ups in the evaluated methods of post-processing.

	ME	MAE	RMSE
ANN 4 hidden neurons	0.8406	1.6856	11.8038
ANN 3 hidden neurons	0.4088	1.8395	11.8919
RLS λ=0.95	0.1203	2.1109	12.3525
RLS λ=1.00	0.0538	2.1911	12.7302
MLR 6 predictors	0.5957	2.1503	13.0064
MLR 3 predictors	1.0369	2.2140	13.4703

The following table shows the same results but using the cross-correlation procedure and Vector Of Displacement approach.

	ME	MAE	RMSE
ANN 6 hidden neurons	0.0036	1.6283	11.5729
ANN 3 hidden neurons	-0.0775	1.6971	11.7552
RLS λ=0.95	1.2364	2.0847	12.1510
RLS λ=1.00	-0.7295	2.1130	12.4476
MLR 6 predictors	0.6641	2.1769	12.9326
MLR 4 predictors	1.2260	2.1990	13.3877

The figures below show exemplary results for the average MAE/RMSE values for Direct Model Output and after using the VOD procedure.

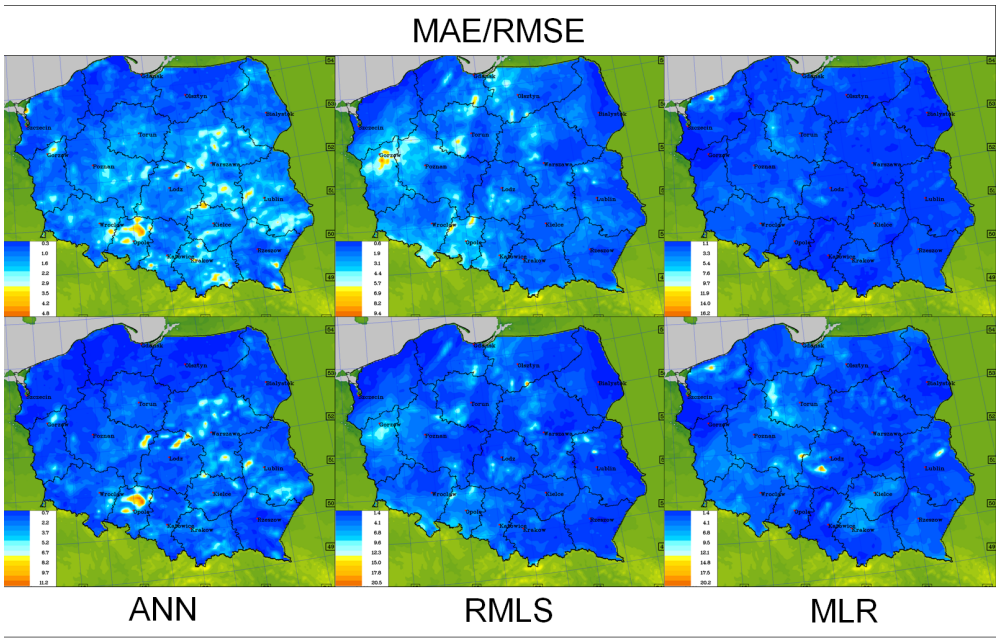


Figure 112: Results for the average MAE/RMSE values for Direct Model Output of Flash Rate (observation vs. forecasts).

Conclusions

Of all the methods, ANN appears to be the best, basing on the results expressed as the MAE and RMSE. This confirms the results that has been already obtained in post-processing with EPS.

When VOD procedure is applied to MAE/RMSE, slight improvement can be seen in comparison to direct verification, with a maxima of MAE/RMSE shifted towards centre of the domain. A similar effect was recognized for all values.

The Recursive/Adaptive Least Mean Square method not necessarily works as good as expected (i.e. the results are not better than the ANN results), but they are much better compared to the standard Multi-Line Regression approach.

Extended works are planned to improve the Flash Rate post-processing methods, however, in the frame of the newly established Priority Project MILEPOST (MachIne Learning-based POST-processing).

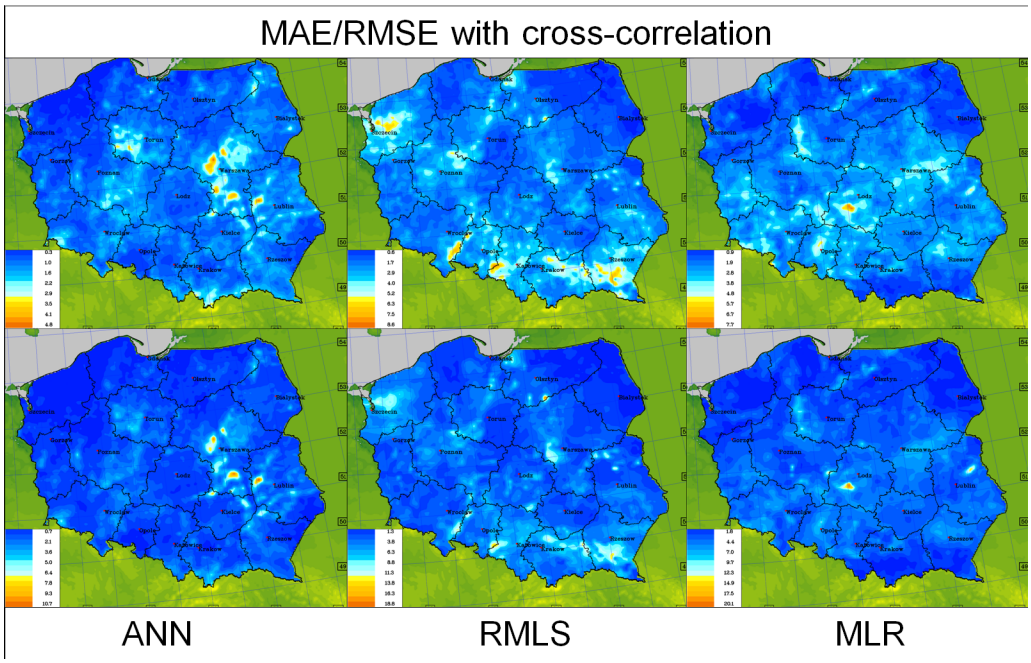


Figure 113: As in Fig. 112, but after using the VOD (cross-correlation) procedure.

7.3 QPF evaluation approaches

ARPAE-SIMC, Maria Stefania Tesini

Introduction

The evaluation of the amount of precipitation over catchment areas is one of the most important uses of the QPF at ARPAE for hydrological purposes and for the issuing of Civil Protection alert for possible floods. To meet the needs of end-users, such as hydrologists or forecasters, some tools that provide mean, maximum and some other percentile values of the precipitation field over the catchment areas of the Emilia-Romagna region have been developed. Exceeding predefined thresholds can give useful indications for situations of intense precipitation possibly leading to floods.

Description of the products

To evaluate the hydrological response of a basin it is not necessary (although desirable) to know precisely the exact location of the amount of precipitation but it is fundamental to have an estimate of the total amount of water that will fall on the area of interest. Results of verification based on the DIST methodology applied to the warning areas (as shown in task 3.4) encourage using the average and the maximum of the precipitation of the points that fall on the area as good products derived from models forecasts.

Each day, Arpae forecasters must provide hydrologists and Civil Protection Department with an assessment of the expected average precipitation on the warning basins based on the data of the models available to them, such as COSMO-5M (5 Km horizontal resolution), COSMO-2I (2.2 Km horizontal resolution) and IFS-ECMWF (9 km horizontal resolution).

Figure 114: Catchment area of Emilia-Romagna region.

To facilitate the comparison of the QPF of these models, summary tables with estimated mean and maximum precipitation over each of the eight catchment areas of the Emilia-Romagna region are produced by means of LIBSIM software developed at Arpae (https://github.com/ARPA-SIMC/libsim).

For each model, it is possible to visualize the estimated average precipitation over each catchment area by step of 6 or 24 hours for the available period of forecast, as shown in figure below. It is also possible to display a text file in which are tabulated the also the mean and the maximum value of the precipitation in each area and the number of points that exceed increasing thresholds (1, 5,10,20,50,100 mm in 6h).

It is also possible to compare models forecast with observed mean values of the previous days using the same tool in order to have a quick validation of the forecast.

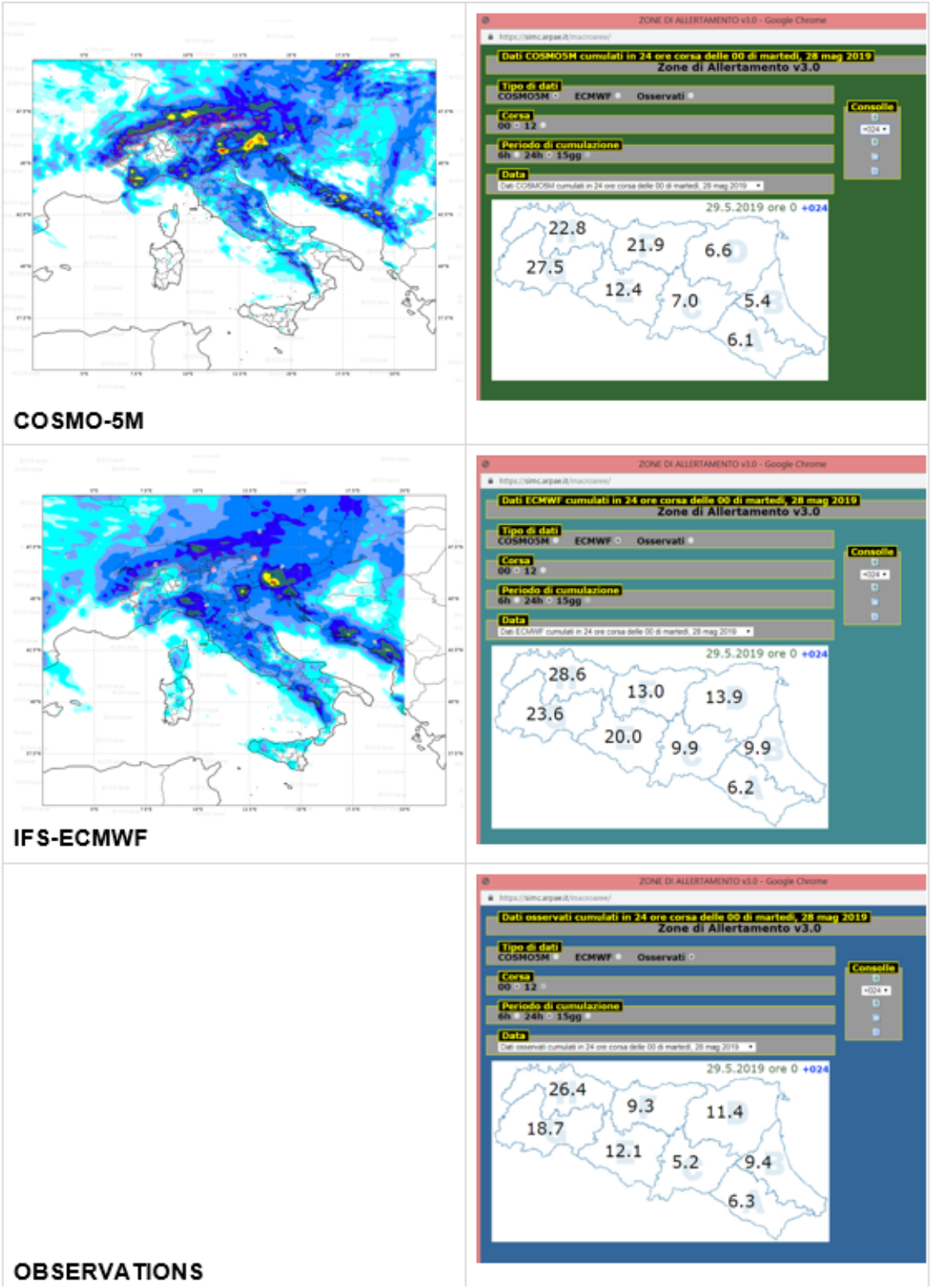


Figure 115: Example of total precipitation field and corresponding average value on Emilia-Romagna catchment areas of COSMO-5M (top) and IFS-ECMWF (middle) and corresponding observations average values (bottom) available the day after.

Ar	Sca	Valida il	alle	Npt	Media	VMedia	Max	Var	>001mm	>005mm	>010mm	>020mm	>050mm	>100mm
1	+024	2019-05-29	00:00	99	6.1	6.1	14.9	3.2	96.0	60.6	7.1	0.0	0.0	0.0
1	+030	2019-05-29	06:00	99	7.1	7.1	17.0	4.2	85.9	70.7	20.2	0.0	0.0	0.0
1	+036	2019-05-29	12:00	99	5.0	5.0	14.9	3.9	79.8	48.5	13.1	0.0	0.0	0.0
1	+042	2019-05-29	18:00	99	4.3	4.3	14.6	3.5	74.7	41.4	7.1	0.0	0.0	0.0
1	+048	2019-05-30	00:00	99	11.0	11.0	58.3	12.5	83.8	56.6	34.3	20.2	2.0	0.0
1	+054	2019-05-30	06:00	99	19.0	19.0	65.1	15.4	100.0	87.9	56.6	38.4	5.1	0.0
1	+060	2019-05-30	12:00	99	17.0	17.0	58.6	13.9	100.0	82.8	54.5	34.3	3.0	0.0
1	+066	2019-05-30	18:00	99	16.5	16.5	58.2	13.9	100.0	80.8	49.5	34.3	3.0	0.0
1	+072	2019-05-31	00:00	99	9.8	9.8	29.3	7.5	100.0	70.7	38.4	11.1	0.0	0.0
2	+024	2019-05-29	00:00	101	5.4	5.4	12.4	2.0	100.0	48.5	4.0	0.0	0.0	0.0
2	+030	2019-05-29	06:00	101	6.2	6.2	12.4	2.2	100.0	68.3	5.9	0.0	0.0	0.0
2	+036	2019-05-29	12:00	101	5.9	5.9	15.7	2.6	100.0	62.4	6.9	0.0	0.0	0.0
2	+042	2019-05-29	18:00	101	3.5	3.5	24.8	3.6	77.2	21.8	3.0	1.0	0.0	0.0
2	+048	2019-05-30	00:00	101	7.9	7.9	53.4	7.8	98.0	59.4	22.8	7.9	1.0	0.0
2	+054	2019-05-30	06:00	101	9.9	9.9	53.1	8.7	97.0	67.3	37.6	12.9	1.0	0.0
2	+060	2019-05-30	12:00	101	9.0	9.0	51.6	7.9	99.0	66.3	26.7	8.9	1.0	0.0
2	+066	2019-05-30	18:00	101	7.8	7.8	51.6	7.8	98.0	48.5	22.8	6.9	1.0	0.0
2	+072	2019-05-31	00:00	101	3.3	3.3	16.9	3.1	85.1	19.8	4.0	0.0	0.0	0.0
3	+024	2019-05-29	00:00	115	7.0	7.0	17.4	4.1	97.4	64.3	21.7	0.0	0.0	0.0
3	+030	2019-05-29	06:00	115	6.7	6.7	15.9	3.9	95.7	61.7	20.0	0.0	0.0	0.0
3	+036	2019-05-29	12:00	115	3.1	3.1	12.0	2.4	81.7	16.5	2.6	0.0	0.0	0.0
3	+042	2019-05-29	18:00	115	5.7	5.7	21.0	5.4	73.9	47.0	17.4	1.7	0.0	0.0
3	+048	2019-05-30	00:00	115	7.4	7.4	19.7	4.4	97.4	63.5	23.5	0.0	0.0	0.0
3	+054	2019-05-30	06:00	115	8.1	8.1	19.3	4.1	99.1	71.3	25.2	0.0	0.0	0.0
3	+060	2019-05-30	12:00	115	8.2	8.2	19.4	4.1	99.1	73.0	27.0	0.0	0.0	0.0
3	+066	2019-05-30	18:00	115	3.1	3.1	9.3	2.0	90.4	23.5	0.0	0.0	0.0	0.0
3	+072	2019-05-31	00:00	115	0.9	0.9	6.0	1.4	21.7	3.5	0.0	0.0	0.0	0.0
4	+024	2019-05-29	00:00	169	6.6	6.6	27.6	5.6	95.3	50.9	23.7	4.1	0.0	0.0
4	+030	2019-05-29	06:00	169	6.0	6.0	29.1	6.0	92.9	38.5	21.3	4.7	0.0	0.0
4	+036	2019-05-29	12:00	169	7.8	7.8	29.1	6.1	94.7	59.2	36.7	5.3	0.0	0.0
4	+042	2019-05-29	18:00	169	14.9	14.9	35.7	8.7	91.1	83.4	68.6	30.8	0.0	0.0
4	+048	2019-05-30	00:00	169	11.5	11.5	31.2	6.8	96.4	84.6	53.3	11.2	0.0	0.0
4	+054	2019-05-30	06:00	169	11.7	11.7	31.2	6.7	100.0	85.2	53.3	11.8	0.0	0.0
4	+060	2019-05-30	12:00	169	10.1	10.1	24.5	5.0	100.0	85.2	45.0	3.6	0.0	0.0

Figure 116: Example of tabular file with number of points exceeding some thresholds and some statistical index for each area and forecast step.

Using the COSMO-LEPS system, we also evaluate the probability of exceeding some thresholds of average precipitation in 24 hours over all the 133 Italian catchment areas.

The product was initially developed for the Emilia-Romagna region as a table in which rows represent the catchment area of the Emilia-Romagna region, columns the threshold (mm/24) and the color of the cell the probability of exceeding the corresponding threshold.

The probability is evaluated considering the average precipitation on the area of interest for each of the members of the ensembles. The product has been subsequently extended to all 133 Italian alert areas with a new graphical version. It is possible to visualize these products for the 5 days of the Cosmo-LEPS forecast. It should be pointed out that we do not use thresholds on probability to issue alert, but they help forecaster to assess confidence in one modeling chain or the other.

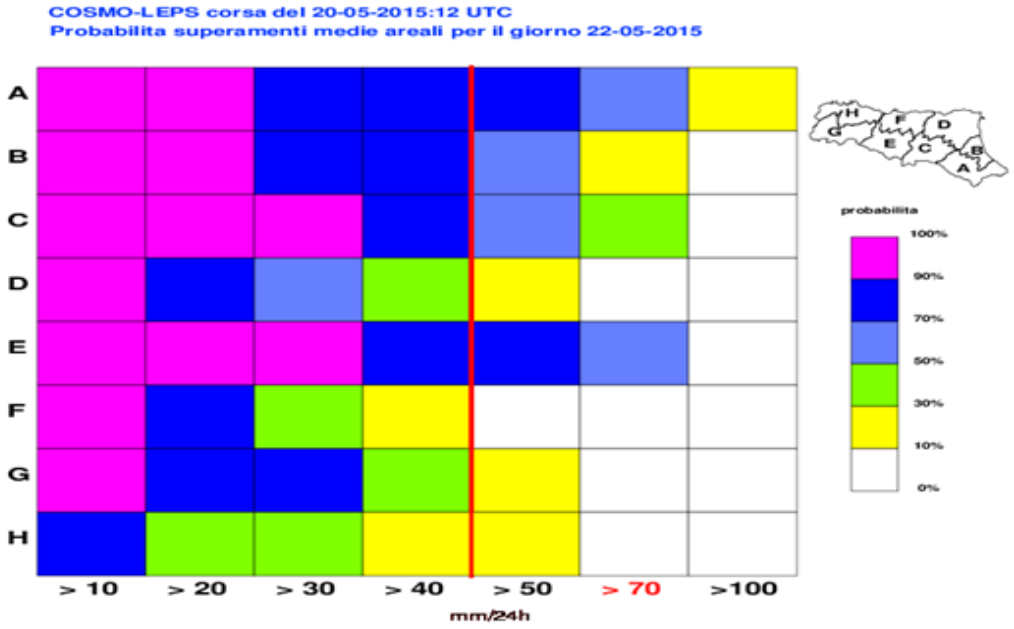


Figure 117: Probability of exceeding increasing thresholds of the average areal precipitation based on the COMSO system (indicated by the colours). In the table rows represent the catchment area of the Emilia-Romagna region, while columns the threshold (mm/24).

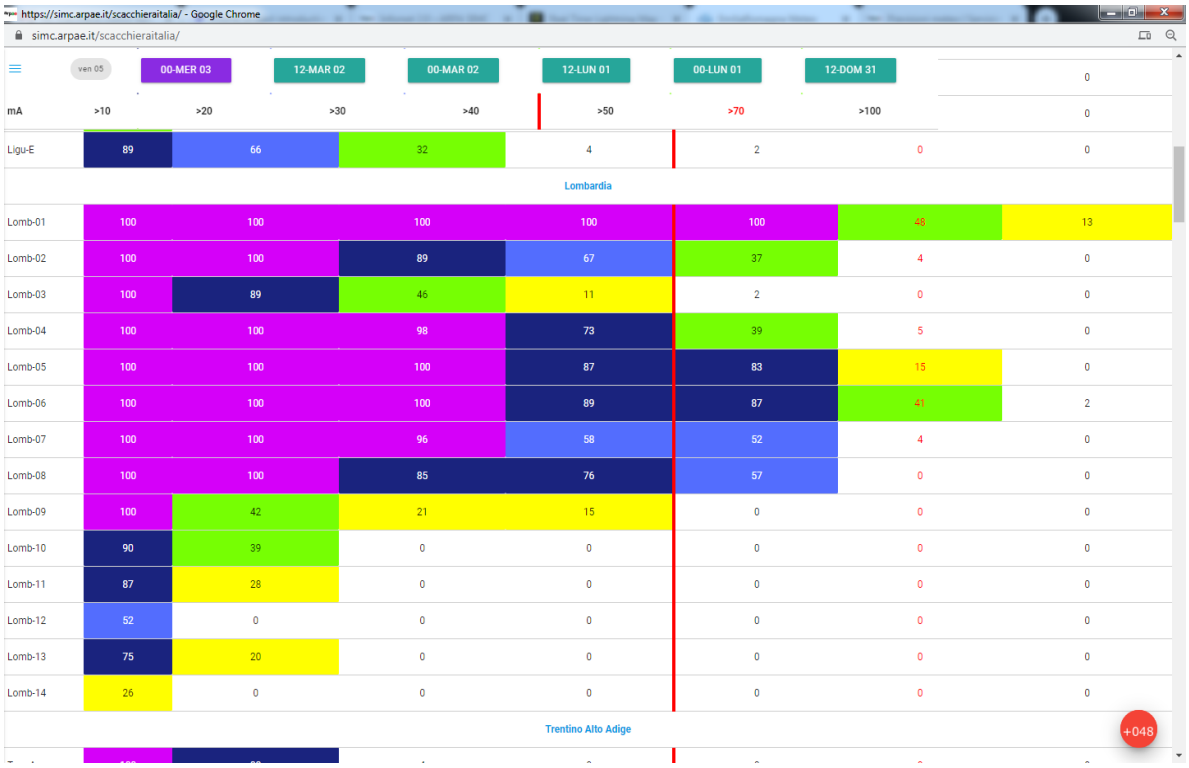


Figure 118: Probability of exceeding increasing thresholds of the average areal precipitation based on the COMSO system (indicated by the colors). The rows represent the catchment areas subdivided by regions, columns represent increasing threshold (mm/24).

Validation

Deterministic products for each warning area are validated on a seasonal basis using "bubbles plot" charts, a sort of the scatter plot in which the data points are replaced with bubbles and the sizes of the bubbles are determined by the number of events. The advantage of this approach is that the nature of the forecast errors can more easily be diagnosed.

Observed and forecast precipitation, aggregated on the catchment areas are divided into classes for average and maximum precipitation on the area and separate plots for each indicator are produced.

CLASSES FOR MEAN PRECIPITATION	MEAN AMOUNT IN 24h (mm)
NO PRECIPITATION	<0.2
NON SIGNIFICANT	0.2 – 5
LIGHT	5-20
MODERATE	20-45
HEAVY	>45

CLASSES FOR MAX PRECIPITATION							
MAX AMOUNT IN 24h (mm)	0.2 -5	5-25	25-50	50-75	75-100	100-150	>150

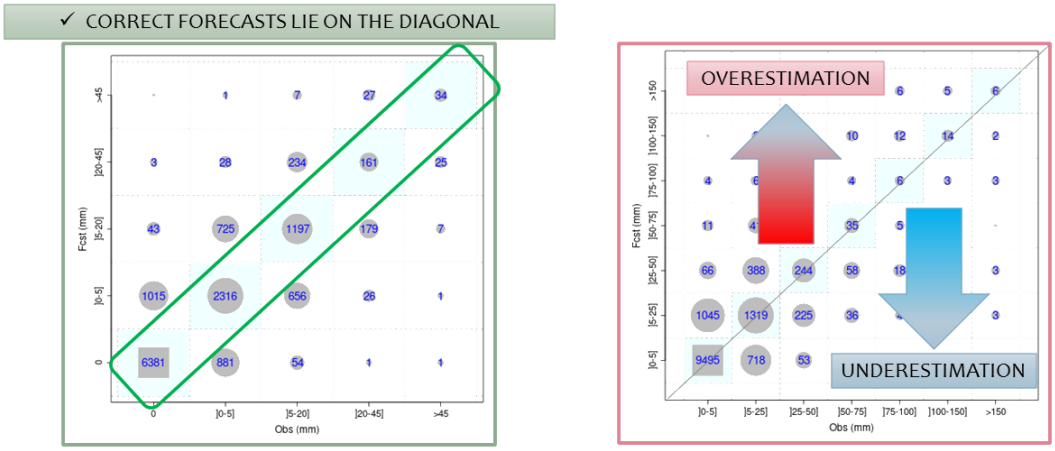


Figure 119: "Bubble plots chart" is a sort scatter plot in which the data points are replaced with bubbles and the sizes of the bubbles are determined by the number of events (when the number of events is large, a square symbol is used for the most populated category to preserve the proportions of the other bubbles). The nature of prediction errors can be easily diagnosed based on the position of the bubbles relative to the diagonal, which represents the correct predictions.

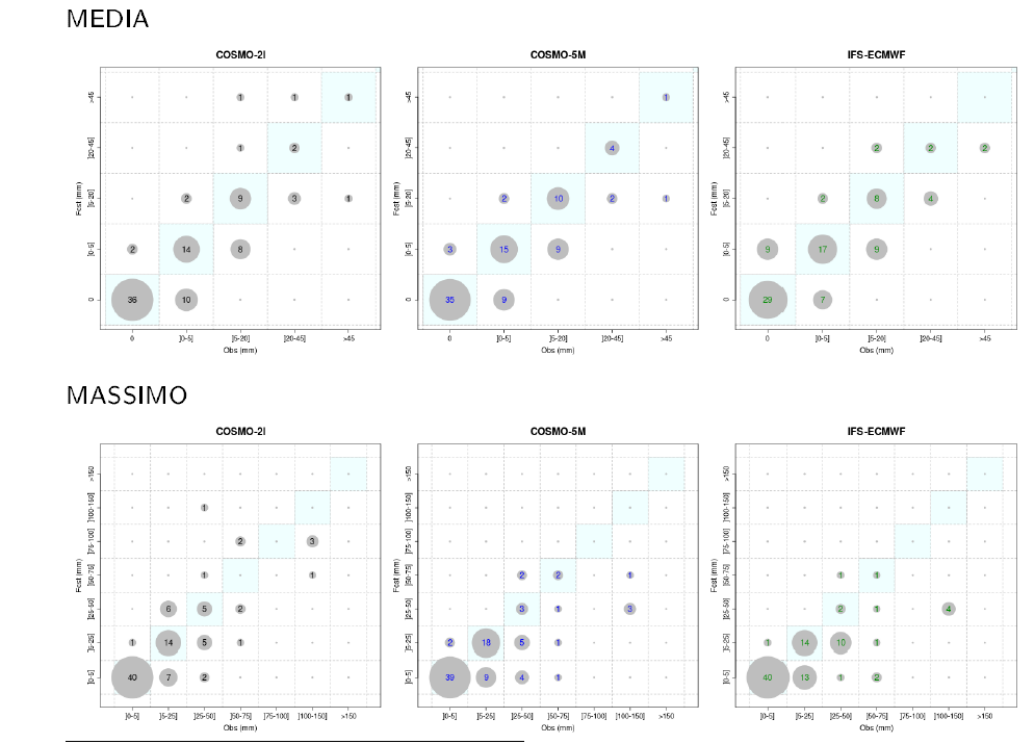


Figure 120: Example of "bubbles plot" relative to an area of the Emilia-Romagna region, as presented in the Arpae seasonal report for MAM2019. In the top panel are displayed the charts for mean value, in the bottom panel those for maximum for the three models (COSMO-2I, COSMO-5M,IFS-ECMWF from left to right).

The validation has been extended to all the Italian catchment areas and reports (in Italian), starting from 2018, they are produced and made available as pdf document to several users (forecasters/hydrologist) on a seasonal basis.

In addition to charts for each single area, summary plots aggregating the data for all the Italian areas are also produced. In the following plots results for the last four seasons are shown.

Figure 121: DJF2019-2020, summary for all Italian catchment areas.

In the top panel are displayed the charts for mean value, in the bottom panel those for maximum for the three models (COSMO-2I, COSMO-5M, IFS-ECMWF from left to right) considering all the Italians catchment areas

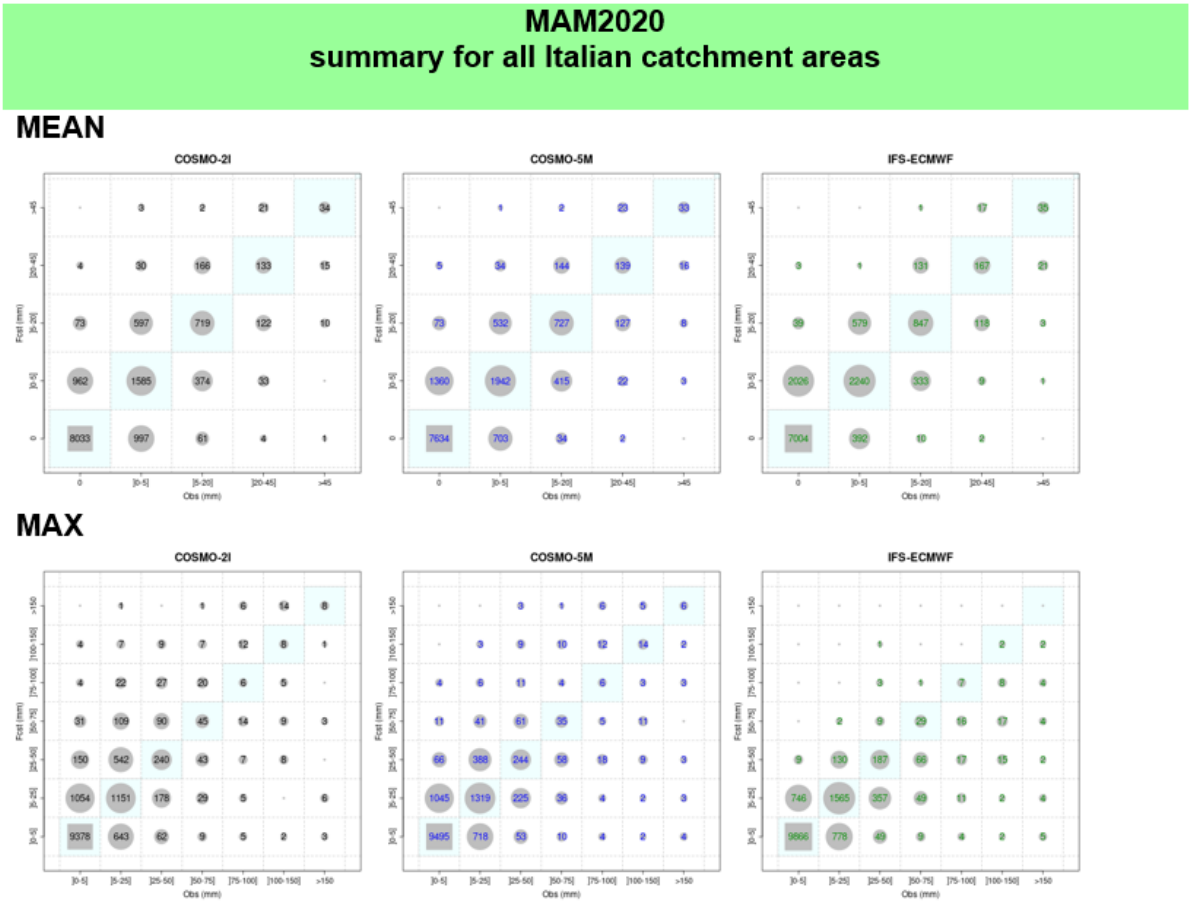


Figure 122: MAM2020, summary for all Italian catchment areas.

In the top panel are displayed the charts for mean value, in the bottom panel those for maximum for the three models (COSMO-2I, COSMO-5M, IFS-ECMWF from left to right) considering all the Italians catchment areas.

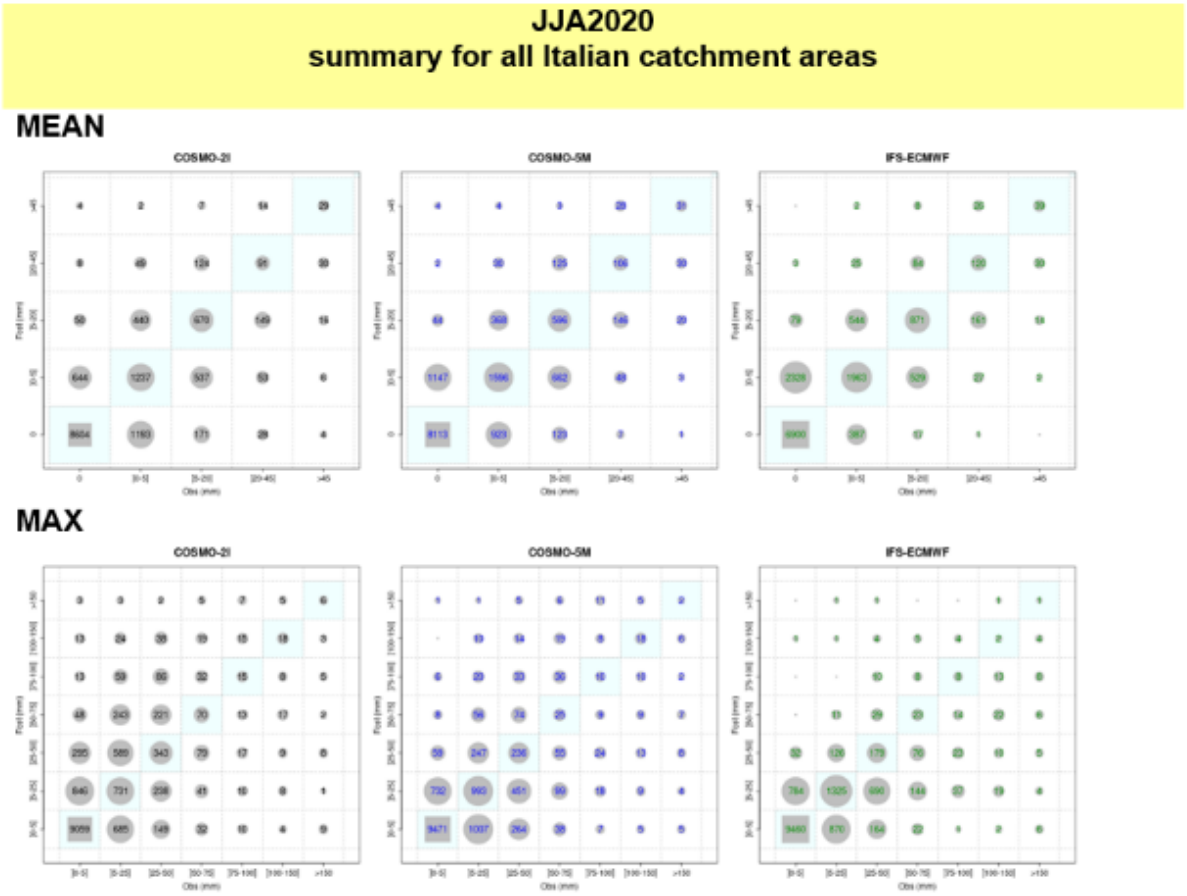


Figure 123: JJA2020, summary for all Italian catchment areas.

In the top panel are displayed the charts for mean value, in the bottom panel those for maximum for the three models (COSMO-2I, COSMO-5M, IFS-ECMWF from left to right) considering all the Italians catchment areas.

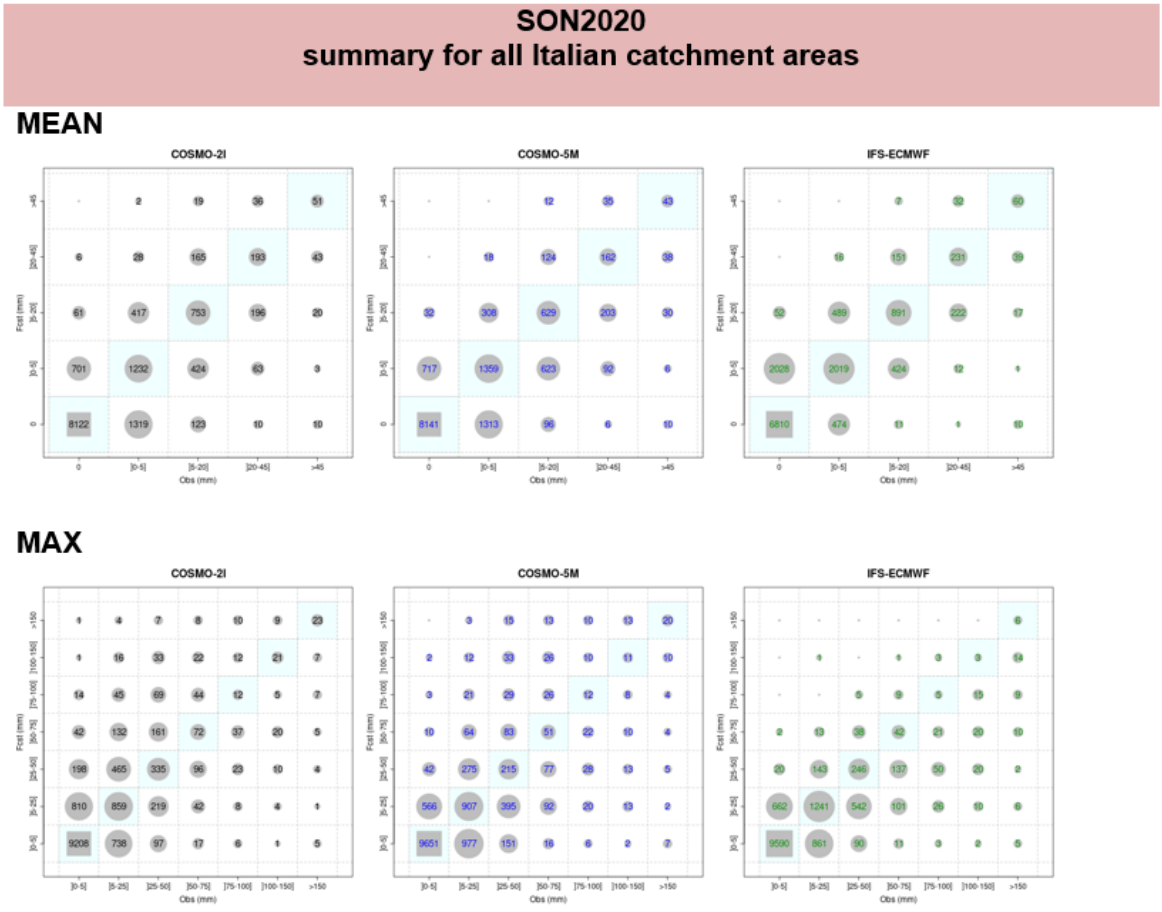


Figure 124: SON2020, summary for all Italian catchment areas.

In the top panel are displayed the charts for mean value, in the bottom panel those for maximum for the three models (COSMO-2I, COSMO-5M, IFS-ECMWF from left to right) considering all the Italians catchment areas.

8 General References

(specific references are given after each chapter in the text)

WMO HIW implementation plan

https://www.wmo.int/pages/prog/arep/wwrp/new/documents/ HIW_IP_v1_4.pdf

High Impact Weather newsletter, September 2018.

https://www.wmo.int/pages/prog/arep/wwrp/new/documents/HIWeather_news_September_2018.pdf

Int. J. Dis. Risk Red. Special Issue Papers

https://www.sciencedirect.com/journal/international-journal-of-disaster-risk-reduction/vol/30/part/PA

Communicating high impact weather: Improving warnings and decision making processes, Andrea Louise Taylor, Thomas Kox, David Johnston, Int. J. Dis. Risk Red. Special Issue Pages 1-4

The influence of impact-based severe weather warnings on risk perceptions and intended protective actions, Sally H. Potter, Peter V. Kreft, Petar Milojev, Chris Noble, ... Sarah Gauden-Ing, Int. J. Dis. Risk Red. Special Issue, Pages 34-43

Towards user-orientated weather warnings, Thomas Kox, Harald Kempf, Catharina Lüder, Renate Hagedorn, Lars Gerhold, Int. J. Dis. Risk Red. Special Issue, Pages 74-80

Bąkowski R., Achimowicz J., Mazur A. (2014): Current Status of Early Warning Systems for Severe Environmental Threats In The Polish National Meteorological Service. Meteorol. Hydrol. Water Manage. 2(2):35–42, DOI: https://doi.org/10.26491/mhwm/36432

Gubenko I., A study the physical processes in convective clouds during thunderstorms based on numerical simulation, PhD thesis, Moscow, 2016 (In Russian).

Losee J. L., Joslyn S. The need to trust: How features of the forecasted weather influence forecast trust, Int. J. Dis. Risk Red. Special Issue, Pages 95-104

Gentine, P., M. Pritchard, S. Rasp, G. Reinaudi and G. Yacalis, 2018: Could machine learning break the convection parametrization deadlock?, *Geophys. Res. Lett.* (Link to online article)

Karstens, C.D.K. et al, 2018, Development of a Human–Machine Mix for Forecasting Severe Convective Events, Mon Wea Rev DOI: 10.1175/WAF-D-17-0188.1

Knox JA, McCann DW, Williams PD. 2008. Application of the Lighthill-Ford theory of spontaneous imbalance to clear-air turbulence forecasting. Journal of Atmospheric Sciences 65: 3292–3304.

Liang,X., et al, 2018, SURF: Understanding and predicting urban convection and haze, BAMS, DOI:10.1175/BAMS-D-16-0178.1

Pantillon, F., Lerch, S., Knippertz, P. and Corsmeier, U.: Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble. Q. J. R. Meteorol. Soc. Accepted Author Manuscript. doi:10.1002/qj.3380. (Link to online article)

Pardowitz, T., 2018, A statistical model to estimate local vulnerability to severe weather. Nat. Haz. Earth Sys.Sci., https://doi.org/10.5194/nhess-18-1617-2018

Radanovics et al, 2018, Spatial Verification of Ensemble Precipitation: An Ensemble Version of SAL, Wea & Forecast, DOI: 10.1175/WAF-D-17-0162.1

Richter H. The severe thunderstorm forecast and warning process in Australia. - Bureau of Meteorology Training centre, Melbourne, Australia, 2008, pp.4-11.

COSMO technical report 37
(http://www.cosmo-model.org/content/model/documentation/techReports/default.htm)

List of COSMO Newsletters and Technical Reports

(available for download from the COSMO Website: www.cosmo-model.org)

COSMO Newsletters

No. 1: February 2001.

No. 2: February 2002.

No. 3: February 2003.

No. 4: February 2004.

No. 5: April 2005.

No. 6: July 2006.

No. 7: April 2008; Proceedings from the 8th COSMO General Meeting in Bucharest, 2006.

No. 8: September 2008; Proceedings from the 9th COSMO General Meeting in Athens, 2007.

No. 9: December 2008.

No. 10: March 2010.

No. 11: April 2011.

No. 12: April 2012.

No. 13: April 2013.

No. 15: July 2015.

No. 16: July 2016.

No. 17: July 2017.

No. 18: November 2018.

No. 19: October 2019.

No. 20: December 2020.

No. 21: May 2022.

No. 22: May 2023.

COSMO Technical Reports

No. 1: Dmitrii Mironov and Matthias Raschendorfer (2001):
Evaluation of Empirical Parameters of the New LM Surface-Layer Parameterization Scheme. Results from Numerical Experiments Including the Soil Moisture Analysis.

No. 2: Reinhold Schrodin and Erdmann Heise (2001):
The Multi-Layer Version of the DWD Soil Model TERRA_LM.

No. 3: Günther Doms (2001):
A Scheme for Monotonic Numerical Diffusion in the LM.

No. 4: Hans-Joachim Herzog, Ursula Schubert, Gerd Vogel, Adelheid Fiedler and Roswitha Kirchner (2002):
LLM — the High-Resolving Nonhydrostatic Simulation Model in the DWD-Project LITFASS.
Part I: Modelling Technique and Simulation Method.

No. 5: Jean-Marie Bettems (2002):
EUCOS Impact Study Using the Limited-Area Non-Hydrostatic NWP Model in Operational Use at MeteoSwiss.

No. 6: Heinz-Werner Bitzer and Jürgen Steppeler (2004):
Documentation of the Z-Coordinate Dynamical Core of LM.

No. 7: Hans-Joachim Herzog, Almut Gassmann (2005):
Lorenz- and Charney-Phillips vertical grid experimentation using a compressible nonhydrostatic toy-model relevant to the fast-mode part of the 'Lokal-Modell'.

No. 8: Chiara Marsigli, Andrea Montani, Tiziana Paccagnella, Davide Sacchetti, André Walser, Marco Arpagaus, Thomas Schumann (2005):
Evaluation of the Performance of the COSMO-LEPS System.

No. 9: Erdmann Heise, Bodo Ritter, Reinhold Schrodin (2006):
Operational Implementation of the Multilayer Soil Model.

No. 10: M.D. Tsyrulnikov (2007):
Is the particle filtering approach appropriate for meso-scale data assimilation ?

No. 11: Dmitrii V. Mironov (2008):
Parameterization of Lakes in Numerical Weather Prediction. Description of a Lake Model.

No. 12: Adriano Raspanti (2009):
COSMO Priority Project "VERification System Unified Survey" (VERSUS): Final Report.

No. 13: Chiara Marsigli (2009):
COSMO Priority Project "Short Range Ensemble Prediction System" (SREPS): Final Report.

No. 14: Michael Baldauf (2009):
COSMO Priority Project "Further Developments of the Runge-Kutta Time Integration Scheme" (RK): Final Report.

No. 15: Silke Dierer (2009):
COSMO Priority Project "Tackle deficiencies in quantitative precipitation forecast" (QPF): Final Report.

No. 16: Pierre Eckert (2009):
COSMO Priority Project "INTERP": Final Report.

No. 17: D. Leuenberger, M. Stoll and A. Roches (2010):
Description of some convective indices implemented in the COSMO model.

No. 18: Daniel Leuenberger (2010):
Statistical analysis of high-resolution COSMO Ensemble forecasts in view of Data Assimilation.

No. 19: A. Montani, D. Cesari, C. Marsigli, T. Paccagnella (2010):
Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO–LEPS system: main achievements and open challenges.

No. 20: A. Roches, O. Fuhrer (2012):
Tracer module in the COSMO model.

No. 21: Michael Baldauf (2013):
A new fast-waves solver for the Runge-Kutta dynamical core.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_21

No. 22: C. Marsigli, T. Diomede, A. Montani, T. Paccagnella, P. Louka, F. Gofa, A. Corigliano (2013):
The CONSENS Priority Project.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_22

No. 23: M. Baldauf, O. Fuhrer, M. J. Kurowski, G. de Morsier, M. Müllner, Z. P. Piotrowski, B. Rosa, P. L. Vitagliano, D. Wójcik, M. Ziemiański (2013):
The COSMO Priority Project 'Conservative Dynamical Core' Final Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_23

No. 24: A. K. Miltenberger, A. Roches, S. Pfahl, H. Wernli (2014):
Online Trajectory Module in COSMO: a short user guide.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_24

No. 25: P. Khain, I. Carmona, A. Voudouri, E. Avgoustoglou, J.-M. Bettems, F. Grazzini (2015):
The Proof of the Parameters Calibration Method: CALMO Progress Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_25

No. 26: D. Mironov, E. Machulskaya, B. Szintai, M. Raschendorfer, V. Perov, M. Chumakov, E. Avgoustoglou (2015):
The COSMO Priority Project 'UTCS' Final Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_26

No. 27: J-M. Bettems (2015):
The COSMO Priority Project 'COLOBOC': Final Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_27

No. 28: Ulrich Blahak (2016):
RADAR_MIE_LM and RADAR_MIELIB - Calculation of Radar Reflectivity from Model Output.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_28

No. 29: M. Tsyrulnikov and D. Gayfulin (2016):
A Stochastic Pattern Generator for ensemble applications.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_29

No. 30: D. Mironov and E. Machulskaya (2017):
A Turbulence Kinetic Energy – Scalar Variance Turbulence Parameterization Scheme.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_30

No. 31: P. Khain, I. Carmona, A. Voudouri, E. Avgoustoglou, J.-M. Bettems, F. Grazzini, P. Kaufmann (2017):
CALMO - Progress Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_31

No. 32: A. Voudouri, P. Khain, I. Carmona, E. Avgoustoglou, J.M. Bettems, F. Grazzini, O. Bellprat, P. Kaufmann and E. Bucchignani (2017):
Calibration of COSMO Model, Priority Project CALMO Final report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_32

No. 33: N. Vela (2017):
VAST 2.0 - User Manual.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_33

No. 34: C. Marsigli, D. Alferov, M. Arpagaus, E. Astakhova, R. Bonanno, G. Duniec, C. Gebhardt, W. Interewicz, N. Loglisci, A. Mazur, V. Maurer, A. Montani, A. Walser (2018):
COsmo Towards Ensembles at the Km-scale IN Our countries (COTEKINO), Priority Project final report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_34

No. 35: G. Rivin, I. Rozinkina, E. Astakhova, A. Montani, D. Alferov, M. Arpagaus, D. Blinov, A. Bundel, M. Chumakov, P. Eckert, A. Euripides, J. Förstner, J. Helmert, E. Kazakova, A. Kirsanov, V. Kopeikin, E. Kukanova, D. Majewski, C. Marsigli, G. de Morsier, A. Muravev, T. Paccagnella, U. Schättler, C. Schraff, M. Shatunova, A. Shcherbakov, P. Steiner, M. Zaichenko (2018):
The COSMO Priority Project CORSO Final Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_35

No. 36: A. Raspanti, A. Celozzi, A. Troisi, A. Vocino, R. Bove, F. Batignani (2018):
The COSMO Priority Project VERSUS2 Final Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_36

No. 37: A. Bundel, F. Gofa, D. Alferov, E. Astakhova, P. Baumann, D. Boucouvala, U. Damrath, P. Eckert, A. Kirsanov, X. Lapillonne, J. Linkowska, C. Marsigli, A. Montani, A. Muraviev, E. Oberto, M.S. Tesini, N. Vela, A. Wyszogrodzki, M. Zaichenko, A. Walser (2019):
The COSMO Priority Project INSPECT Final Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_37

No. 38: G. Rivin, I. Rozinkina, E. Astakhova, A. Montani, J-M. Bettems, D. Alferov, D. Blinov, P. Eckert, A. Euripides, J. Helmert, M.Shatunova (2019):
The COSMO Priority Project CORSO-A Final Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_38

No. 39: C. Marsigli, D. Alferov, E. Astakhova, G. Duniec, D. Gayfulin, C. Gebhardt, W. Interewicz, N. Loglisci, F. Marcucci, A. Mazur, A. Montani, M. Tsyrulnikov, A. Walser (2019):
Studying perturbations for the representation of modeling uncertainties in Ensemble development (SPRED Priority Project): Final Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_39

No. 40: E. Bucchignani, P. Mercogliano, V. Garbero, M. Milelli, M. Varentsov, I. Rozinkina, G. Rivin, D. Blinov, A. Kirsanov, H. Wouters, J.-P. Schulz, U. Schättler (2019):
Analysis and Evaluation of TERRA_URB Scheme: PT AEVUS Final Report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_40

No. 41: X. Lapillonne, O. Fuhrer (2020):
Performance On Massively Parallel Architectures (POMPA): Final report.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_41

No. 42: E. Avgoustoglou, A. Voudouri, I Carmona, E. Bucchignani, Y. Levy, J. -M. Bettems (2020):
A methodology towards the hierarchy of COSMO parameter calibration tests via the domain sensitivity over the Mediterranean area.
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_42

No. 43: H. Muskatel, U. Blahak, P. Khain, A. Shtivelman, M. Raschendorfer, M. Kohler, D. Rieger, O. Fuhrer, X. Lapillonne, G. Rivin, N. Chubarova, M. Shatunova, A. Poliukhov, A. Kirsanov, T. Andreadis, S. Gruber (2021):
The COSMO Priority Project $T^2(RC)^2$: Testing and Tuning of Revised Cloud Radiation Coupling, Final Report
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_43

No. 44: M. Baldauf, D. Wojcik, F. Prill, D. Reinert, R. Dumitrache, A. Iriza, G. deMorsier, M. Shatunova, G. Zaengl, U. Schaettler (2021):
The COSMO Priority Project CDIC: Comparison of the dynamical cores of ICON and COSMO, Final Report
DOI: 10.5676/DWD_pub/nwv/cosmo-tr_44

No. 45 Marsigli C., Astakhova E. Duniec G., Fuezer L., Gayfulin D., Gebhardt C., Golino R., Heppelmann T., Interewicz W., Marcucci F., Mazur A., Sprengel M., Tsyrulnikov M., Walser A. (2022):
The COSMO Priority Project APSU: Final Report.

No. 46 A. Iriza-Burca, F. Gofa, D. Boucouvala, T. Andreadis, J. Linkowska, P. Khain, A. Shtivelman, F. Batignani, A. Pauling, A. Kirsanov, T. Gastaldo, B. Maco, M. Bogdan, F. Fundel (2022):
The COSMO Priority Project CARMA: Common Area with Rfdbk/MEC Application Final Report.

No. 47 A. Voudouri, E. Avgoustoglou, Y. Levy, I. Carmona, E. Bucchignani, J. M. Bettems (2022):
Calibration of COSMO Model, Priority Project CALMO-MAX: Final Report.

No. 48 D. Rieger et al. (2022):
The Priority Project C2I, Transition of COSMO to ICON - Final Report.

No. 49 E. Churiulin, M. Toelle, V. Kopeikin, M. Uebel, J. Helmert and J.-M. Bettems (2022):
The COSMO Priority Task VAINT: Vegetation Atmosphere INTeractions Report.

COSMO Technical Reports

Issues of the COSMO Technical Reports series are published by the *COnsortium for Small-scale MOdelling* at non-regular intervals. COSMO is a European group for numerical weather prediction with participating meteorological services from Germany (DWD, AWGeophys), Greece (HNMS), Italy (USAM, ARPA-SIMC, ARPA Piemonte), Switzerland (MeteoSwiss), Poland (IMGW), Romania (NMA) and Russia (RHM). The general goal is to develop, improve and maintain a non-hydrostatic limited area modelling system to be used for both operational and research applications by the members of COSMO. This system is initially based on the COSMO-Model (previously known as LM) of DWD with its corresponding data assimilation system.

The Technical Reports are intended

- for scientific contributions and a documentation of research activities,
- to present and discuss results obtained from the model system,
- to present and discuss verification results and interpretation methods,
- for a documentation of technical changes to the model system,
- to give an overview of new components of the model system.

The purpose of these reports is to communicate results, changes and progress related to the LM model system relatively fast within the COSMO consortium, and also to inform other NWP groups on our current research activities. In this way the discussion on a specific topic can be stimulated at an early stage. In order to publish a report very soon after the completion of the manuscript, we have decided to omit a thorough reviewing procedure and only a rough check is done by the editors and a third reviewer. We apologize for typographical and other errors or inconsistencies which may still be present.

At present, the Technical Reports are available for download from the COSMO web site (www.cosmo-model.org). If required, the member meteorological centres can produce hard-copies by their own for distribution within their service. All members of the consortium will be informed about new issues by email.

For any comments and questions, please contact the editor:

Massimo Milelli

massimo.milelli@cimafoundation.org