Survey for assessment of proper verification of phenomena

Andrzej Mazur, Joanna Linkowska

Institute of Meteorology and Water Management – National Research Institute

Report on sub-task 2.1 of COSMO Priority Project AWARE (Appraisal of Challenging WeAther FoREcasts)

Contents

i. 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,
- The 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:

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

ii. Methods

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

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

where all the above further on called as “discrete” analysis

- **Standard evaluation at the grid scale**

hereinafter referred to as “continuous” analysis

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

[1] 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

[2] 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

Structure-Amplitude-Location (SAL) analysis.

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

- **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.

- **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.

- **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.

The most common case is marked with bold. As it can be seen the parameterization of flashrate intensity based on the CAPE (cf. Report on task 3.1 Priority Project AWARE) generally overestimates FR compared to the observations.

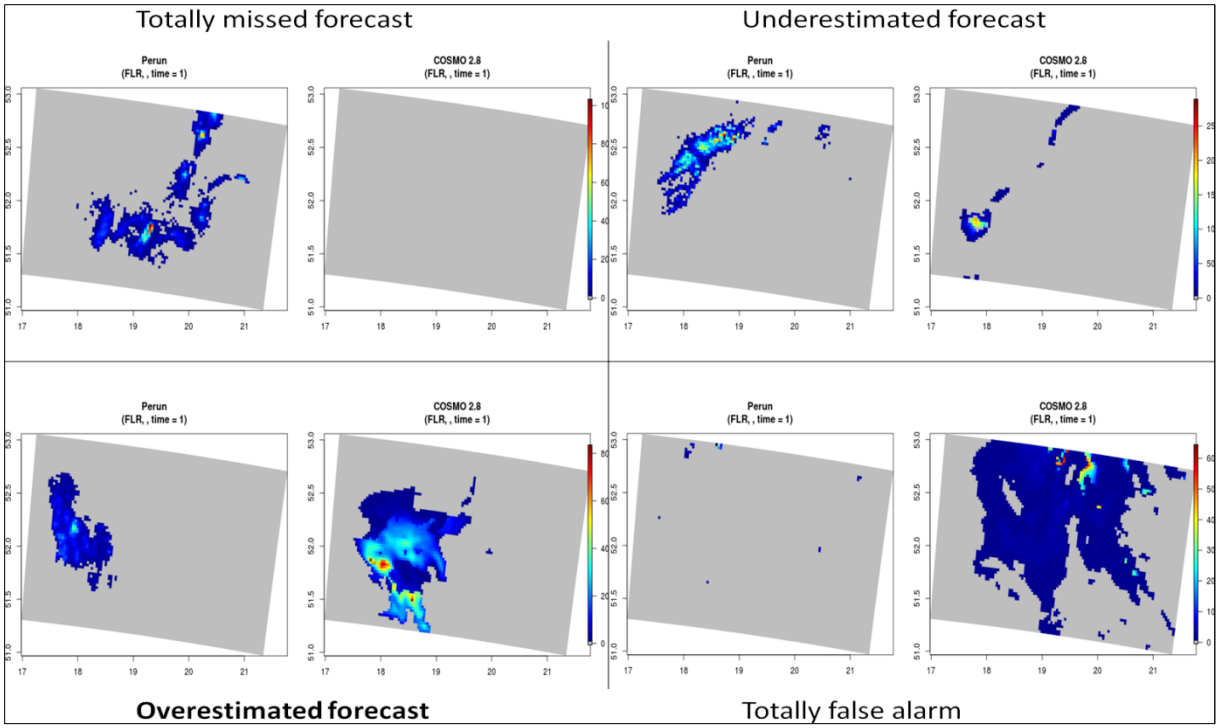

Fig. 1 Exemplary verification of flashrate intensity forecasts – Structure-Amplitude-Location approach.

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).

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.

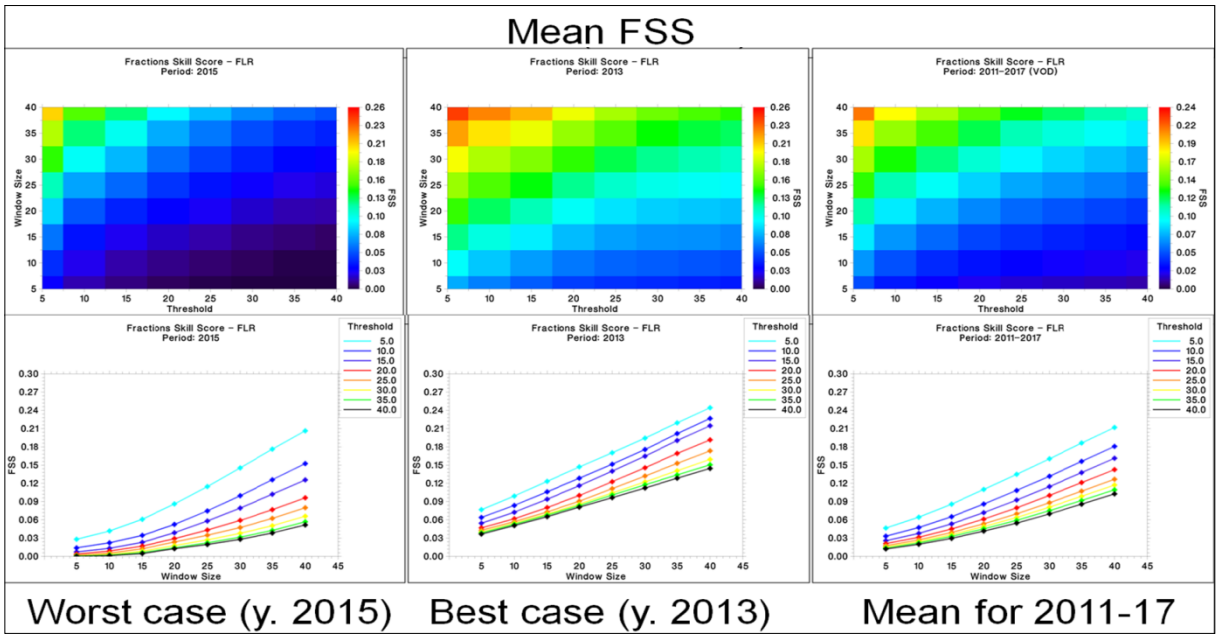

Fig. 2 Results of FSS for the worst (2015), the best (2013) year and mean for the entire period of 2011-2017; parameterization of flashrate intensity based on the CAPE (*ibidem*).

Categorical analysis based on contingency tables

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	$\frac{a \cdot d - b \cdot c}{(a+c) \cdot (b+d)}$	-1 to 1	1
Equitable Skill Score	$\frac{a - a_r}{(a+b+c-a_r)}$ $a_r = \frac{(a+b) \cdot (a+c)}{n}$	-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. 3.

	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

Tab. 1 Results of contingency tables analysis for the entire period of 2011-2017; parameterization of flashrate intensity based on the CAPE (*ibidem*).

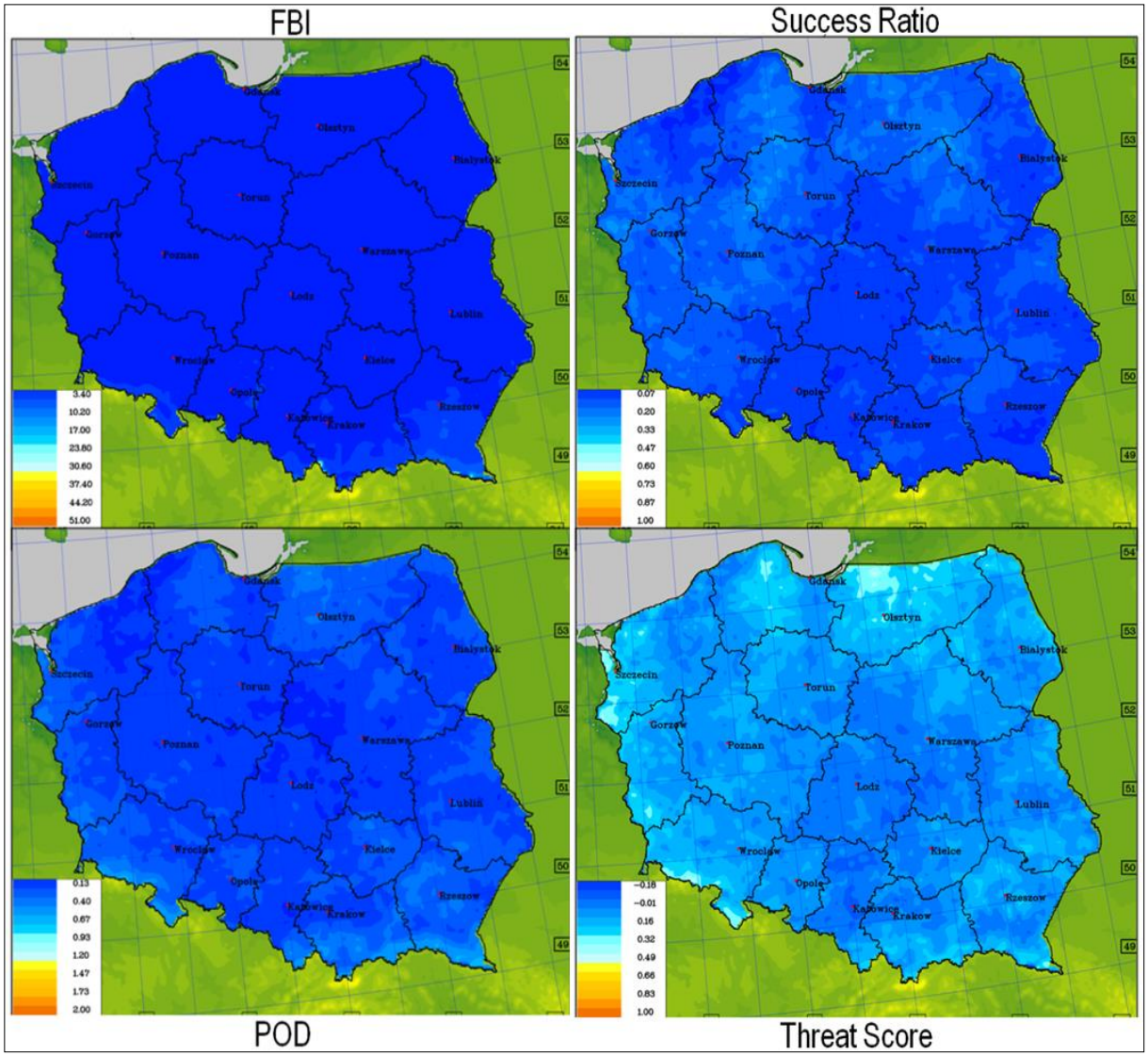

Fig. 3 Results of contingency tables analysis for the entire period of 2011-2017 – selected predictands; parameterization of flashrate intensity based on the CAPE (*ibidem*).

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

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.

Tab. 2. ME/MAE/RMSE for consecutive years and mean values for 2011-2017; parameterization of flashrate intensity based on the CAPE (*ibidem*)

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

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

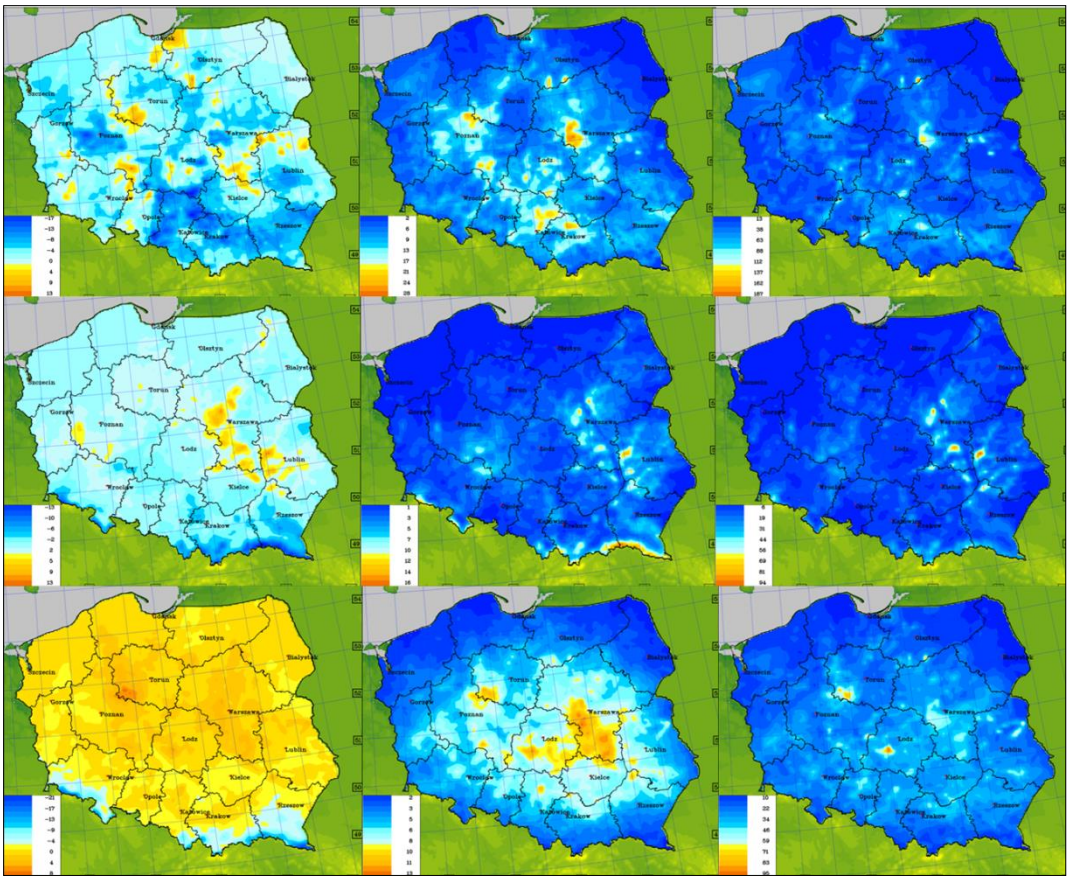

Fig. 4 Left to right: ME, MAE and RMSE for 2013, 2017 and mean 2013-2017 as in table 2.

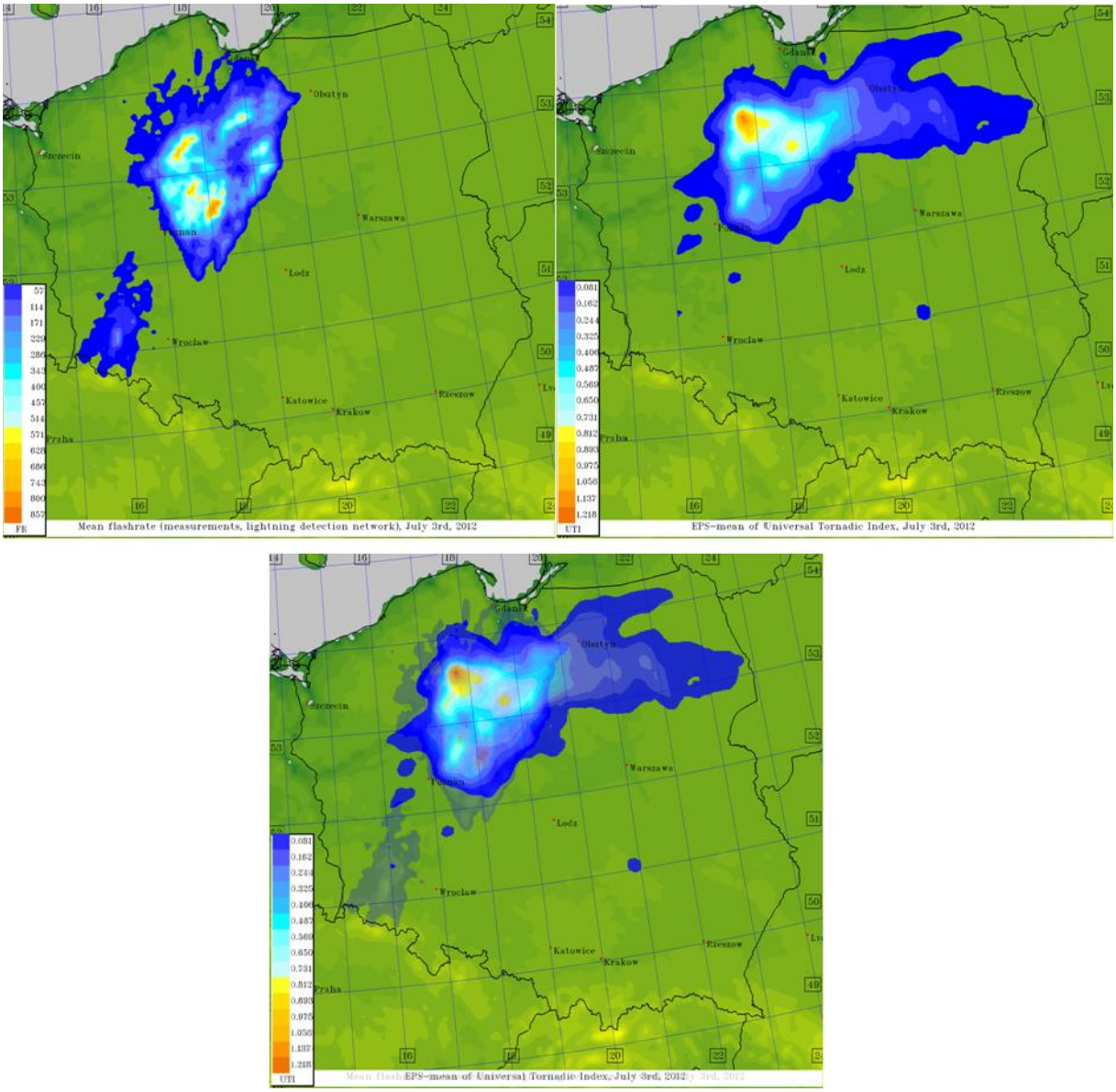

Fig. 5 Explanation of VOD procedure – see details in text.

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:

- Calculate coordinates of ”centers of mass” for both distribution patterns (observations vs. forecasts).
- Compute vector of displacement (VOD) of forecasts to observations as a difference of the two above.
- 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.

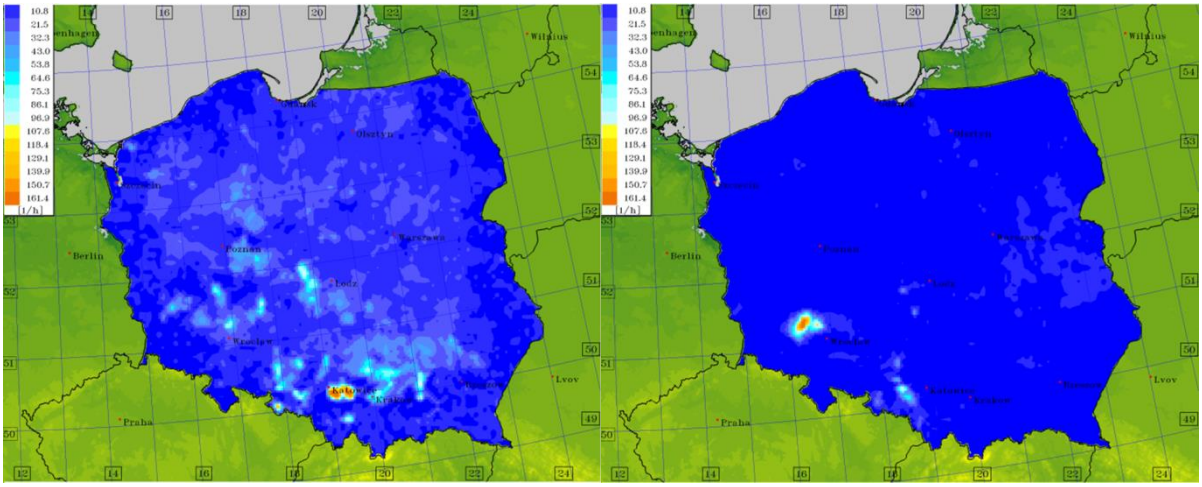

Fig. 6. Sample values of (observations – forecasts) for flash rate (lightning frequency). Left - direct model output results, right panel - corrected with VOD procedure.

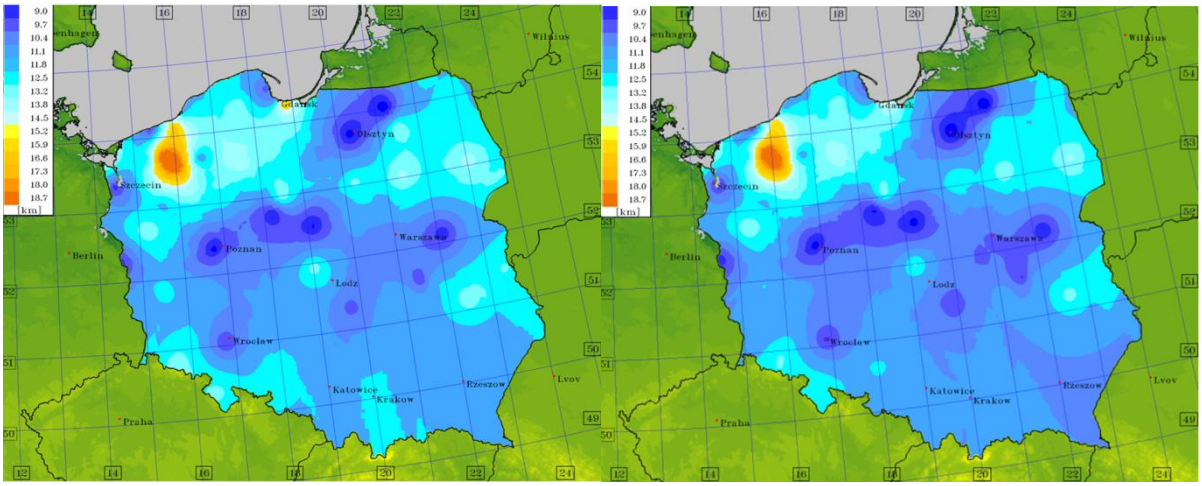

Fig. 7. Sample values of (observations – forecasts) for visibility range. Left - direct model output results, right panel - corrected VOD procedure.

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% (*ibidem*).

iii. 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.

iv. Conclusions

In the next part of the report (subtask 3.1), 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.