Bulgarian Journal of Meteorology and Hydrology

Bulgarian Academy of Sciences

National Institute of Meteorology and Hydrology

Bul. J. Meteo & Hydro 22/3-4 (2017) 73-86

The Operative System ProData–Part One: Current Stage and Recent Improvements

Hristo Chervenkov*, Valery Spiridonov, Eram Artinyan, Plamen Neytchev, Kiril Slavov, Minka Stoyanova

National Institute of Meteorology and Hydrology- BAS,
Tsarigradsko shose 66, 1784 Sofia, Bulgaria

Abstract. The present article, which is part one of several common works, describes the operative system ProData—the reasons and motivation for its creation, the embedded processing technique, the input/output data flow, as well as its strengths. The possibility of the system to adopt further improvements and some achievements in this direction are described and visualized. The main conclusion is that the system is a reliable source of consistent meteorological information with high spatial and temporal resolution with minimal latency from the input data acquisition time.

Keywords: Operative System, Meteorological Data Processing, Automatic Weather Station, Satellite-derived Data, SWEEP Operator

1. INTRODUCTION

The modern applied meteorology is faced with the challenge of the growing demand on reliable data, available in high resolution, both in time and space. Numerous mesoscale geophysical tasks and applications need such data: practically all spatially distributed hydrological and ecological models need certain meteorological information, most frequently formatted as initial data set, containing the values of some input parameters. Thus, for instance, they use air temperature to drive processes such as evapotranspiration, snowmelt, soil water and temperature evolution, and plant productivity. As fundamental meteorological variable, rainfall is primary input for hydrological models, specifically distributed hydrological ones. The regional climate and weather prediction models also rely on such data for verification and tuning. In the common case all near-surface

* hristo.tchervenkov@meteo.bg

weather observations are collected at irregularly spaced point locations (for example the network of the measurement stations) rather than over continuous surfaces. Although the synoptic records are considered to be relatively accurate and reliable at the point where the station locates, the density of the measurement network and frequency of observations are generally not high enough to describe the spatial and temporal distribution of the considered meteorological variables. Often, in many environmental studies, it is impossible to use directly such information or this can lead to serious biases in the results. Consequently, different methods for estimation of spatially and temporally distributed near-surface meteorological variables are developed. They can be pure mathematical, for example inverse-distance weighting (IDW), kriging, 2-dimensional splines, and trend-surface regression (Myers, 1994) or, alternatively, combined-mathematical based on physical assumptions (Chervenkov, 2016). As stated in Dodson and Marks, 1997, these methods often work well over relatively flat, homogeneous terrain. The weather conditions in local scales, however, are partially influenced by the topography of the area. Extensive research was carried out worldwide, partly using the modern GIS technologies, aiming at the accurate visualization and digitization of various climate variables (see Feidas et al 2014 for detailed review). Many efforts are dedicated on developing of appropriate methods for estimating climatic elements using topographical and geographical parameters as independent variables. Such models are able to estimate climate variables in sites that observational data are not available, giving a relatively reliable solution to the old problem of insufficient climatic data. Common weakness of many of these, product of purely geographical approach, is the utilization of only topographical parameters as regressors. Most of the existing solutions work as climate hindcast, i.e. usually the output is produced months after the data acquisition time. Overall, the incorporated techniques are quite sophisticated, the implementation of such methods demands significant computational power and increased amount of input data, making the overall procedure quite difficult. In the group of products from such systems it is worthy to outline the Pan-European gridded dataset on daily basis E-OBS of the European Climate Assessment&Dataset (ECA&D). This dataset is periodically updated, well known in the meteorological community, and widely used for many tasks, extensively as reference in model verification studies (Haylock et al, 2008). As will be commented in the next section, however, the need for operative work of the system is a very significant constraint and has to be always kept in mind.

The paper is structured as follows: Hence this is the first publication, dedicated on ProData, the general description of the system is placed in second section. The most significant recent improvements are concisely reported in the third and fourth section. The main conclusions as well as short outlook of the planned future work are described in the conclusion.

2. SHORT DESCRIPTION OF THE CURRENT STAGE OF PRODATA

The operative system (OS) ProData was created in NIMH-BAS in the period 2012-2015 by the team under the leadership of prof. V. Spiridonov. The basic concept was to combine in methodologically consistent way all available on hourly basis meteorological and auxiliary data in order to produce high-quality gridded time-series, of the most significant meteorological variables. These time-series, or digital maps, have to be with 1 hour time resolution and at least national coverage. The horizontal resolution has to be also adequate, which, according to the modern requirements for the considered scale, is below 10 km. Thus, the current implementation of the system runs on grid with 0.045°×0.045° spacing, which corresponds approximately of 4 km×4 km. The model domain, which is shown on Figure 1, covers entirely Bulgaria and consists of 147×73=10731 grid-cells.

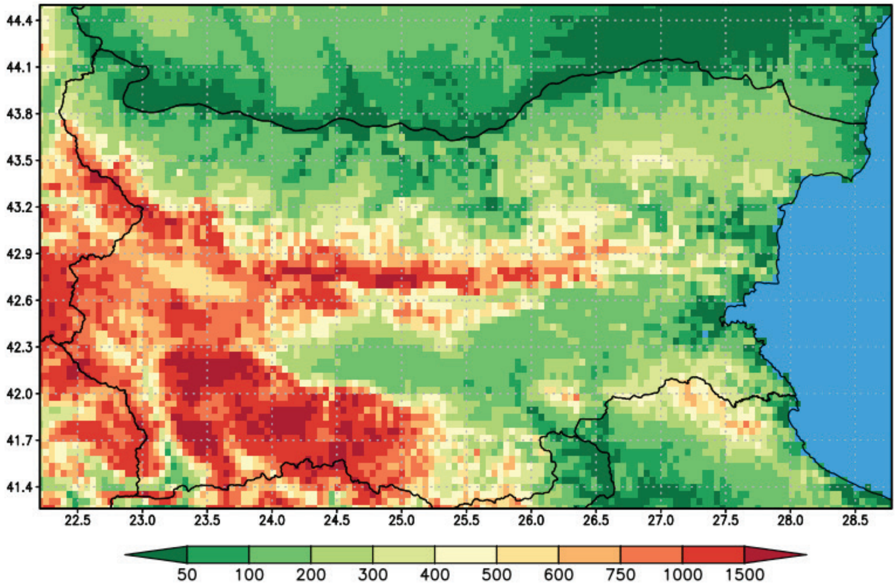


Fig. 1. Model domain and elevation (unit: m) of the grid-cells over land

The specific need of various scientific and other experts was additional motivation. Thus, for example, the operative hydrology needs such data for calculation of the total precipitation amount over a certain river basin; the electricity companies used it for evaluation of the energy consumption, and many more. The requirement for one hour time resolution is a very serious constraint - it narrows significantly the number of potential input data sources. Neither of the traditional (i.e. synoptic and climatic) observational networks measures the meteorological variables in intervals shorter than 3 hours, our radar is currently also disabled. It is necessary in such situation to rely on data, collected and transmitted from *in situ* platforms for environmental monitoring, as automatic weather and hydrological stations (AWS/AHS) as well as satellite-derived data, mainly from the services of the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT).

The network of AWS and AHS was built gradually following the needs of particular projects directed primarily toward streamflow analysis and forecasting as "Flood forecasting and early warning system for Maritsa and Tundzha rivers" (Roelevink et al., 2010), "Flood warning system in Arda river basin - Ardaforecast" (Artinyan et al., 2016), "Danube Water" (Nedkov et al., 2015), etc. The automatic stations measuring hourly precipitation rate over the country are about 140 but only 80 of them have also combined air temperature and relative humidity sensors and 40 of them have solar radiation sensors. These stations are spread irregularly over the country as the above projects covered partially Southern Bulgaria but didn't cover Black Sea river basins for instance. Stations data is collected at hourly basis (Naldzhiyan, 2017) and is exported as ASCII files to be used from ProData system.

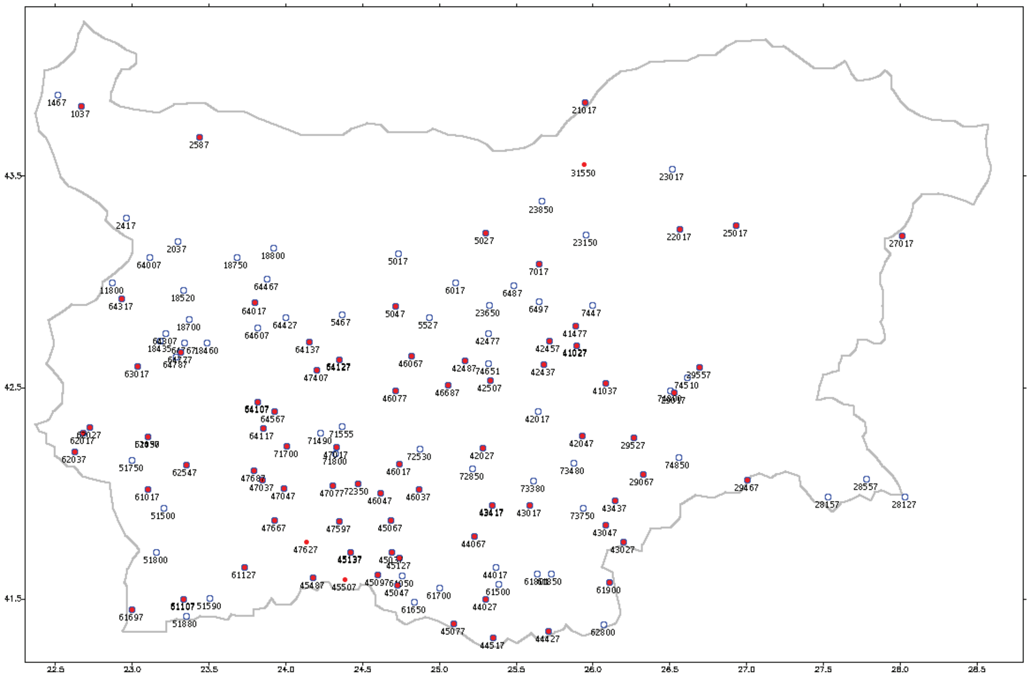


Fig. 2. Network of AWS/AHS, measuring the temperature and relative humidity at 2 m above the ground (red dots) and precipitation amount (blue circles) at every hour.

Being the Bulgarian National Hydromet Service, NIMH-BAS uses widely and decades-long the services and products of EUMETSAT. Since 2014, however, as Bulgaria became a member of the organization, the quantity and variety of the available information, both raw and processed data, have increased remarkably. Consequently, the possibilities for implementation of such data have risen significantly. As will be commented further, the satellite retrievals, more specially these in the infrared (IR) channel of the Meteosat second generation (MSG), are the primary source for the input information for ProData. Currently, gridded estimates for the following 11 variables are routinely produced in the frame of the system every hour:

- temperature and relative humidity at 2m above the surface
- precipitation amount
- cloud coverage
- wind speed and direction
- solar shortwave incoming radiation downwards
- presence of fog, precipitation of hail and thunderstorm
- snow water equivalent (SWE)

Hence the analyzed variables are fairly different and each one has specifics, which have to be taken into account, there is no common procedure for the processing of the input data and, respectively, the preparation of the final product. The wind speed and direction are directly taken from the output of the ALADIN-BG (Bubnova et al., 1995), which is the Bulgarian short-range operational weather forecast model. Many issues have to be addressed intending to combine the information from the geostationary satellite and the network of AWS/AHS. Most of the problems are rooted in the principle differences of the two observation concepts, respectively platforms. Thus, for example, the satellite data have to undergo georeferencing, consequently mapped onto the grid of the system, the error, caused from the parallax and synchronization have also to be taken into account. All corresponding procedures inherently introduce biases, the cumulative effect of which leads to unavoidable limitation of the final product accuracy (for more details see http://www.hydro.bg/mapValej/metodika_za_satelitni_nazemni.pdf).

The core of the system is the objective analysis of the variables. It is done by statistical means, using the well-known multiple linear regression technique (MLR, see appendix). It is widely and successfully implemented in geophysics, partly in climatology (see Feidas et al., 2014 again). The independent variables used are functions of the brightness temperature and also derivatives of the topography, which are calculated prior the MLR. Elevation, exposure, and convexity, which are proportional to the Laplacian of the elevation, belong to the last group. The conceptual scheme and the stepwise data flow processing within ProData are shown on Figure 3. Data from AWS/AHS are used as dependent variables; i. e. the model is forced to mimic the spatial distribution as established from these data. The residual value is the bias between the AWS/AHS-measurements and the final output value. Some authors, in an attempt to refine the estimated values, propose residual correction using different local interpolation methods. The validity of these models is checked through cross-validation error statistics against an independent (test) subset of station data. The benefit of such second step is often questionable: Feidas et al. (2014) finds that the correction of the developed regression models using residuals improved though not significantly the interpolation accuracy.

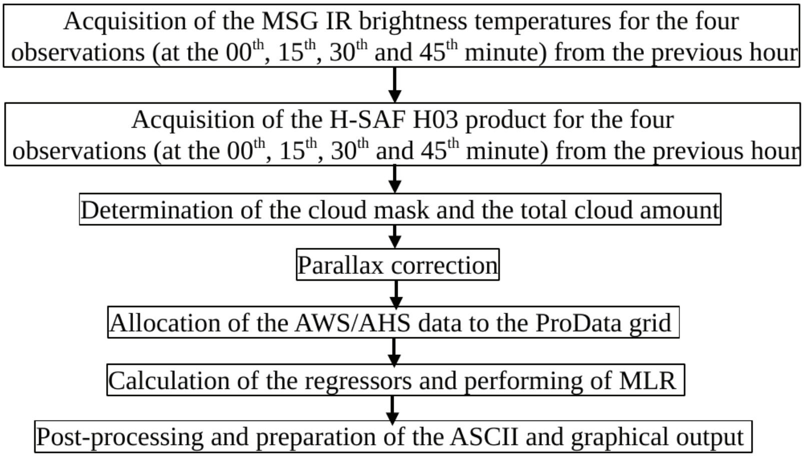

Acquisition of the MSG IR brightness temperatures for the four observations (at the 00th, 15th, 30th and 45th minute) from the previous hour
Acquisition of the H-SAF H03 product for the four observations (at the 00th, 15th, 30th and 45th minute) from the previous hour
Determination of the cloud mask and the total cloud amount
Parallax correction
Allocation of the AWS/AHS data to the ProData grid
Calculation of the regressors and performing of MLR
Post-processing and preparation of the ASCII and graphical output

Fig.3. Data flow and processing steps within ProData

The modules for estimation of fog presence, precipitation of hail and thunderstorm, are similar. All of them use the threshold approach, calculating in advance some criterion quantities. The fog, hail and thunderstorms are determined in two degrees of likelihood of their occurrence depending on the number of conditions fulfilled. For fog the satellite information used follows the criteria, proposed by Barbosa (2012) modified for the Bulgarian conditions. Additional criteria are limitations on wind speed, air temperature, dew point and rainfall. The probability of hail follows the criteria described in Siewert et al. (2010). In the snow module, the snow is accumulated from the precipitation by negative temperatures. The quantitative description of snow pack evolution, including the snow depth determination and SWE, is performed using the methodology, described in the Engineer Manual of the U.S. Army Corps of Engineers (see references).

Concluding this section, we would like to emphasize the main features of the system ProData, on which its success is based:

- The system works in operational (i.e. with minimal latency from the data acquisition time, as a rule approximately 1 hour and 30 minutes) and fully automatic (i.e. unattended from personnel) regime.
- A native computational procedure, coded by the ProData-team, which relies on efficient and transparent statistical technique
- Freely accessible (from within NIMH-BAS private network) trough a web-page.

The basic output products, hourly data sets of all 11 analyzed variables for each grid cell, are available at https://users.meteo.bg/ProData/ in convenient ASCII csv-type format. This site is designed as single point access - it contains also many secondary products, tailored for the specific needs of the different end-users as well as explanatory descriptions and auxiliary data.

3. SOME RECENT IMPROVEMENTS OF THE SYSTEM PRODATA

Significant merit of the system is its flexibility, expressed mainly in the possibility for further enhancement and development. As far as the mathematical approach seems a reasonable choice, it appears to be most perspective to experiment with, and eventually to adopt, new sources of input data as independent variables in the MLR.

A vast quantity of high-quality and reliable environmental data is exchanged nowadays trough the scientific networks or is free-of-charge for non-commercial use. It is expected that this stream, due to the implementation of new methods and platforms on one hand and the increased international cooperation on the other, will rise steadily. This fact is a favorable prerequisite for such experiments. As emphasized before, the requirement for the data acquisition frequency and the transmission latency appear as a principle constraint. Some of the products of the eight satellite application facilities (SAFs) of the EUMETSAT seem promising and especially these, which are directly linked with some of the analyzed within ProData parameters. So far, we have performed an extensive test with the H-SAF PR OBS 3-H03 product. It is worth to emphasize, however, that the other precipitation-linked H-SAF products, despite their advantages over H03, are not suitable due to the timelines constraint: All others are with latency significantly longer than the one-hour limitation.

The primary goal of the Satellite Application Facility in Support to Operational Hydrology and Water Management (H-SAF) is to provide satellite-derived products from existing and future satellites with sufficient time and space resolution to satisfy the needs of operational hydrology. Five of the H-SAF operational products are targeted to the precipitation, and due to the minimal latency, the H-SAF PR OBS 3-H03 one, as stated before, is the single one suitable. Core of the product is the “Rapid Update” technique, which allows computing of instantaneous rain intensities at the ground at the geostationary time-space scale (Turk et al. 2000). It is based on a blended micro-wave (MW)-IR technique that correlates, by means of the statistical probability matching, to brightness temperatures measured by the IR geostationary sensors and passive MW-estimated precipitation rates at the ground. Hence the method suffers from many issues (see the listed in the references Product User Manual for details), this product cannot be used as dependent variable in the MLR, as the data from the AWS/AHS.

The raw ProData output can be preprocessed in order to respond more adequately to the specific needs of the end-users. As a result, various secondary, both numerical and graphical, products could be offered. The biggest share of the ProData web-page content consists already of such data. Maps of the most analyzed parameters in 3-hour interval from 12 UTC of the previous day until the current hour are available on-line on http://hydro.bg. Our intent is to enrich this approach, proposing new figures. The leading idea is to combine optimally clarity and information in as small as possible number of new items. Thus, the collated maps, shown on Figure 4, are specially tailored for quick-view of the meteorological situation in the previous day.

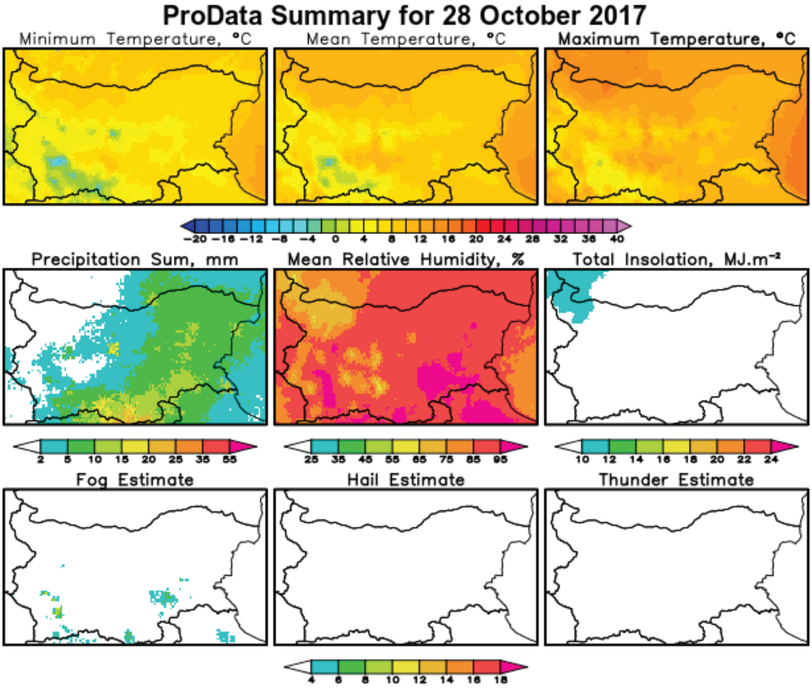
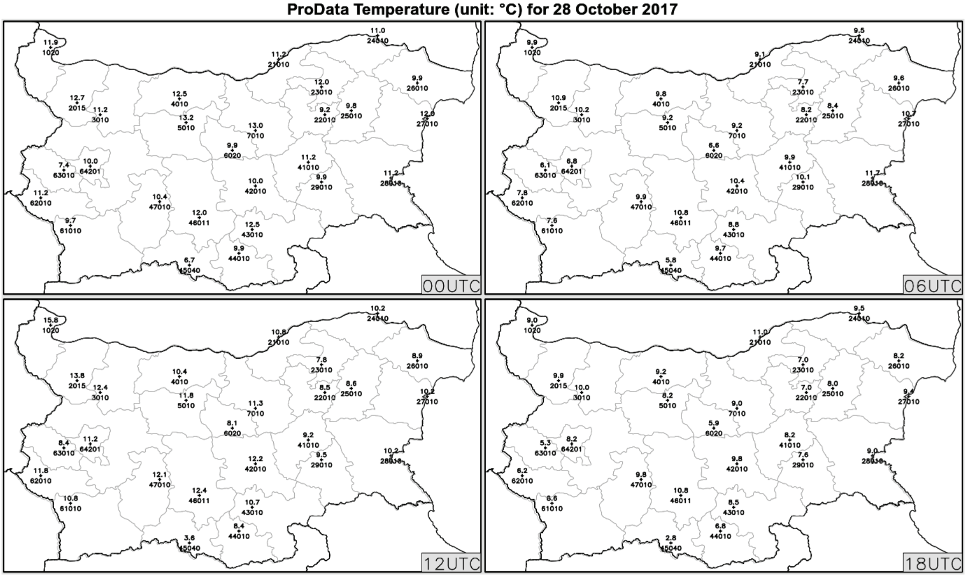


Fig. 4. Quick-view of the meteorological situation

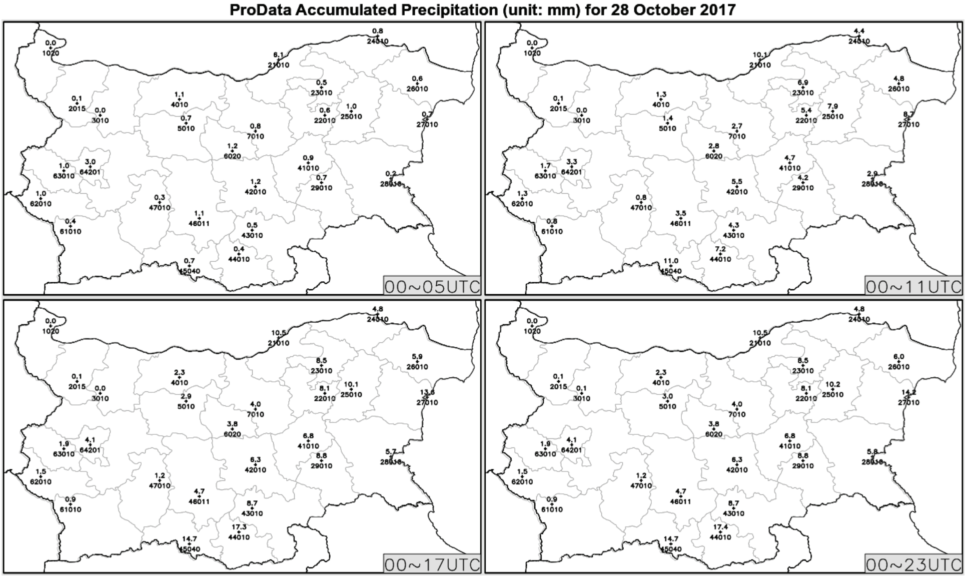
It is worth to emphasize, that the daily minimum, mean, and maximum temperature are calculated a posteriori, using the hourly values of the temperature. These three parameters, together with the daily precipitation amount are most frequently used for estimation of climate extremes and, respectively, they form the standard input data set for the calculation of the climate indices (see, for example, the STARDEX project https://crudata.uea.ac.uk/projects/stardex/)

In some cases and for certain users, as for example representatives of national and municipal authorities, figures with the values in concrete points (i. e. grid cells) of interest, are more suitable rather than color-coded maps. Such figures, for the temperature in the main synoptic terms and the precipitation for four equal intervals, are already automatically generated. Thus far, data for the 27 province centers (i.e. first level administrative subdivisions of the country) are plotted. The possibility however to change easily the list of the considered places, without modification of the main procedure, is already foreseen. Examples of such figures are shown on Figure 5 and Figure 6.



ProData Temperature (unit: °C) for 28 October 2017

Fig. 5. Temperature in the 27 province centers



ProData Accumulated Precipitation (unit: mm) for 28 October 2017

Fig. 6. Precipitation amount in the 27 province centers

The system ProData is also a very convenient source for synoptical and climatological analysis in retrospective manner, including for conducting of hindcast studies. Maps of the day-by-day mean temperature and precipitation amount, as shown on Figure 7 and Figure 8, are very useful in this sense.

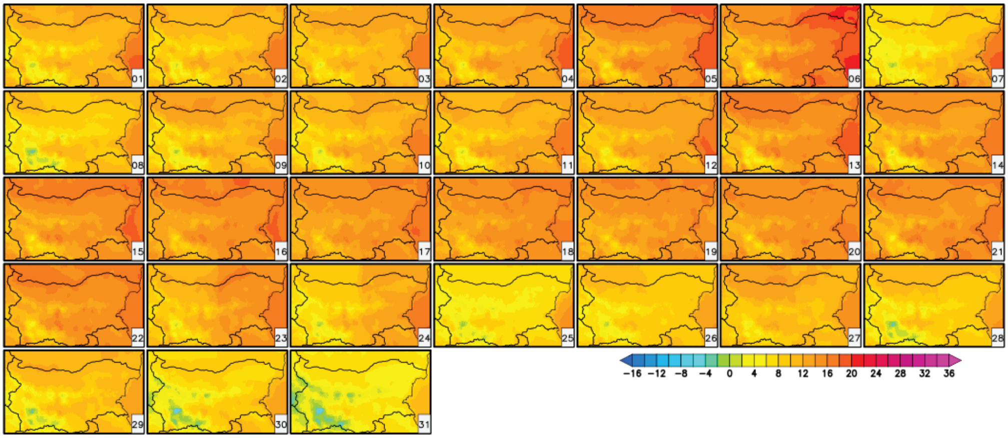


Fig. 7. Daily average temperature for October 2017 (unit: °C)

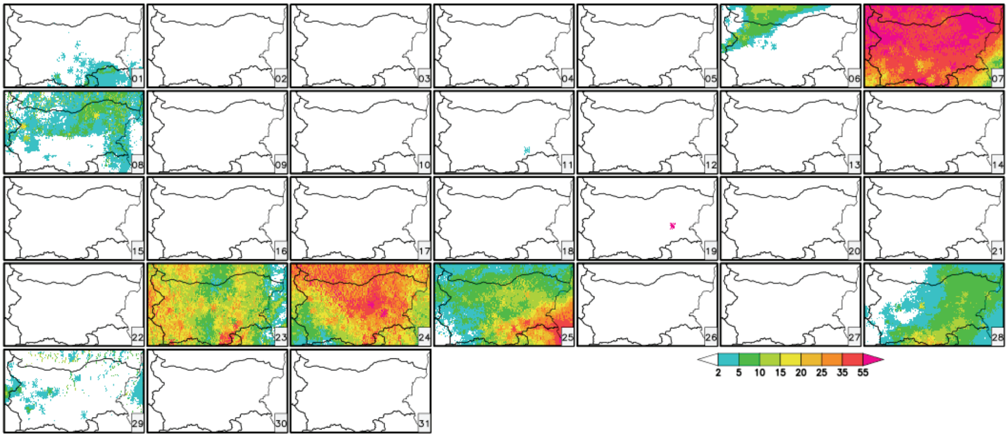


Fig. 8. Daily precipitation amount October 2017 (unit: mm)

A key point in many hindcast studies is to assess the dynamics of the meteorological situation for a certain period of interest. The well-known in the community Grid Analysis and Display System (GrADS), which is used as main graphical pre-processor, provides rich set of built-in functions for spatial and temporal analysis. Thus, with GrADS it is easy to estimate and plot the areal average (AA) of a certain variable as shown on Figure 9.

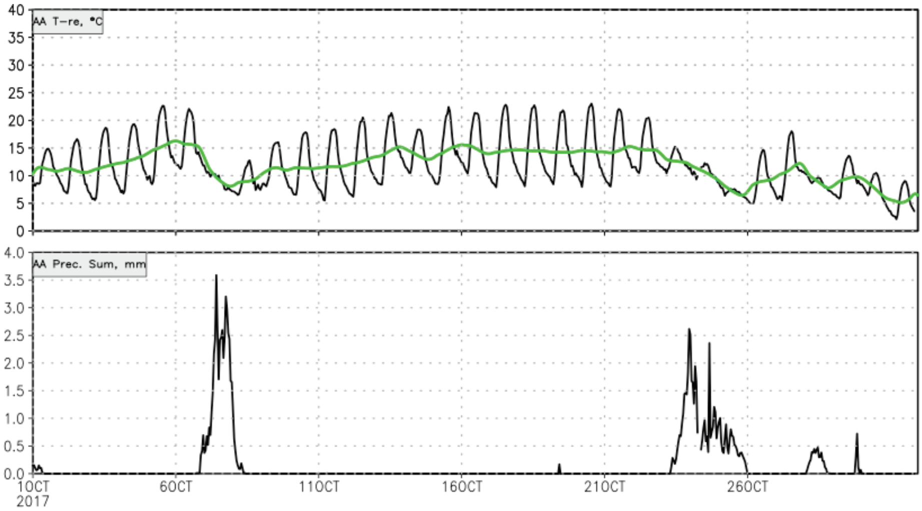


Fig. 9. Monthly chronograms of the AA temperature and precipitation for October 2017

Hence the AA characterizes the domain as a whole, its evolution, unlike the change in a single grid-cell, can be caused only by mesoscale or synoptic reasons. This can be used, in particular, for quick detection of fronts, as illustrated on figure 9: the rapid decrease of the temperature around the 7th October in conjunction with the heavy precipitations over the whole domain, together with the corresponding subplots on figures 8 and 7, suggest passing of a cold front.

It is obvious that the type and quantity of such secondary products can be extended practically with no limits. From at least technical point of view, however, it is reasonable to keep this in certain limits. The authors of the system remain open for any feedback and advice from the community of the end-users.

CONCLUSION

Combining methodological consistency, easy maintenance, transparency and last but not least quick availability of plenty of output data-sets and products, the operative system ProData proves itself as a reliable source of high-quality meteorological information. It is designed as convenient versatile for all, who need single point access of meteorological data in operational mode. Thus, it is used extensively in NIMH-BAS for hydrological short-range forecasts, eventually issuing of warnings. It could be used also in many nowcasting routines for weather forecast activities. Last but not least, ProData is proven to be very robust - practically all cases of failure could be explained with data transfer issues, which are caused by communication problems outside the system. ProData fills in the gap of information in this time and spatial scale and satisfies the needs of various end-users and experts. The necessary next step in our work is to perform in-depth comparison with independent data, which could be treated as reference. This is expected to be the subject of the second part of this article.

ACKNOWLEDGEMENTS

As far as the study is based on free of charge data and open source software, the authors would like to express their deep gratitude to all institutions (EUMETSAT, H-SAF, Unidata, and all others) and individual enthusiasts. This study would not be possible without their innovative data services and tools.

APPENDIX

Multiple linear regression (MLR) is natural generalization of the simple linear one-dimensional model in case of more than one independent variables. MLR attempts to model the relationship between two or more explanatory variables ("regressors") and a response variable by fitting a linear equation to observed data. Formally, the model for MLR, given *n* observations, is:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon_i \quad \text{for} \quad i = 1,2,\ldots n, \tag{1}$$

where $x_i, i = 1,2,\ldots p$ are the independent variables. The term *ε* is called a disturbance term or error variable - an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors. The $p+1$ parameters $\beta_i, i = 0,1,\ldots p$ are referred to as partial regression coefficients, which have to be estimated. Equation (1), which is a system of *n* equations for $p+1$ unknown coefficients, can be rewritten in matrix form as follows:

$$y = X\beta + \varepsilon, \tag{2}$$

where X is a $n \times p$ matrix of the explanatory variables, y is a $n \times 1$ vector of the observations and *β* is a $p \times 1$ vector of the unknown parameters to be estimated. As far as $p+1 < n,$ the linear system in Eq.(2) is overdetermined (i.e. more constraints than variables). Ordinary least squares (OLS) is the simplest and thus most common estimator. It is conceptually simple and computationally straightforward. OLS minimizes the error $S(\hat{\beta}) = \left\| y - X\hat{\beta} \right\|^2$ in meeting the constraint and leads to:

$$\beta = (X^T X)^{-1} X^T y, \tag{3}$$

where β are the estimated regression coefficients. The linear system in Eq. (3) can be solved by means of different methods (e.g. QR- or Cholesky decomposition) including the SWEEP operator as shown in Goodnight (1979) and Neytchev (1995).

REFERENCES

Artinyan, E. et al., (2016): Flood forecasting and alert system for Arda River basin. Journal of Hydrology, 2016, ISSN: 0022-1694

Barbosa, H. (2012) Usefulness of METEOSAT-9 information for fog top detection in Brazil: first measurements, Proceedings of WMO/WWRP International Symposium on Nowcasting and Very Short Range Forecasting

Bubnová, R., Hello, G., Bénard, P., & Geleyn, J. F. (1995). Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the ARPEGE/Aladin NWP system. Monthly Weather Review, 123(2), 515-535

Chervenkov, H. (2016) Simple Postprocessing Method for Vertical Correction of Stratified Near-surface Atmospheric Parameters. Bulgarian Geophysical Journal, 40, ISSN:1311-753X, 14-22

Dodson, R., Marks, D., (1997) Daily air temperature interpolated at high spatial resolution over a large mountainous region Clim Res Vol. 8: 1-20.

Engineer Manual 1110-2-1406, Department of the U.S. Army Corps of Engineers. Washington DC, 20314-1000 RUNOFF FROM SNOWMELT, March 1998. (available on-line at http://www.publications.usace.army.mil/Portals/76/Publications/EngineerManuals/EM_1110-2-1406.pdf?ver=2013-09-04-070756-610)

Goodnight, J. H. (1979) A Tutorial on the SWEEP Operator The American Statistician Vol. 33, No. 3 (Aug., 1979), pp. 149-158

Feidas, H., Karagiannidis, A., Keppas, S. et al. Theor Appl Climatol (2014) 118: 133. https://doi.org/10.1007/s00704-013-1052-4

Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New (2008), A European daily high-resolution gridded dataset of surface temperature and precipitation for 1950 – 2006, J. Geophys. Res., 113, D20119, doi:10.1029/2008JD010201

Myers DE (1994) Spatial interpolation. An overview. Geoderma 62(1):17-28

Naldzhyan A., Georguiev O., Artinyan E., (2017): From the sensors to the models, integrated hydro-meteorological systems in NIMH – BAS, Bulgaria, International Conference on Automatic Weather Stations (ICAWS-2017), 24 – 26 October 2017, Offenbach am Main, Germany

Nedkov, N., et al., (2015): NIMH BG PP10 contribution for the BG - RO common water monitoring and flow forecasting in the CBC region. On-line report. http://danube-water.eu/wp-content/uploads/2015/09/NIMH_4.pdf

Neytchev, Pl. (1995) SWEEP operator for least-squares subject to linear constraints, Computational Statistics & Data Analysis, Vol. 20, Issue 6, 1995, pp. 599-609, ISSN 0167-9473, https://doi.org/10.1016/0167-9473(94)00067-8.

Product User Manual-PUM – 03A (Product H03A-PR-OBS-3A) Doc.No: SAF/HSAF/PUM-03A Issue/Revision Index: 1.2 Date: 10/04 /2015 Page: 1/19

available on-line at http://hsaf.meteoam.it/documents/PUM/SAF_HSAF_PUM-03A_1_2.pdf

Roelevink A., Udo J., Koshinchanov G., Balabanova S., (2010): Flood forecasting system for the Maritsa and Tundzha Rivers, Proceedings of BALWOIS (2010) – Ohrid, Republic of Macedonia – 25, 29 May 2010

Siewert, C. W., Koenig M., Mecikalski J. R. (2010) Application of Meteosat second generation data towards improving the nowcasting of convective initiation, Meteorol. Appl.,17 pp. 442 - 451 DOI: 10.1002/met.176.

Turk J.F., G. Rohaly, J. Hawkins, E.A. Smith, F.S. Marzano, A. Mugnai and V. Levizzani, (2000): “Analysis and assimilation of rainfall from blended SSMI, TRMM and geostationary satellite data”. Proc. 10th AMS Conf. Sat. Meteor. and Ocean., 9, 66-69.