Use of Consolidation Technology for Meteorological Data Processing

Dmitry A. Petrov[1] and Elena N. Stankova[1,2]

[1] Saint-Petersburg State University
198504 St.-Petersburg, Russia, Peterhof, Universitetsky pr., 35
[2] Saint-Petersburg Electrotechnical University "LETI",
197376, St.-Petersburg, Russia, ul.Professora Popova 5
g_q_w_petrov_dm_alex@mail.ru, lena@csa.ru

Abstract. This paper is concerned with the technology of meteorological data consolidation. The technology is used to create one of the components of the virtual environment for forecasting dangerous convective phenomena - thunderstorms, squalls, hail and heavy rainfall. Nowadays, progress in the field of such phenomena simulation is significantly associated with the verification of the already existing models rather than with the development of the new ones [1]. Verification and adjustment of the models are very difficult due to the lack of freely available integrated data on the location and time of the observed phenomena in conjunction with meteorological data used as initial and boundary conditions for numerical models of convective clouds. In the paper we apply consolidation technology to develop the system for heterogeneous data extraction, transformation and loading to the relational database that contains the whole set of meteorological information about the state of the atmosphere at the place and at the time when a dangerous convective phenomenon is recorded. Data sources which are freely available via the Internet are used. The format of the information stored in the database does not require further decoding and can be directly used for the numerical simulation.

Keywords: data consolidation, data integration, meteorological data, numerical modeling, weather forecast.

1 Introduction

The problem of the use of modern numerical models for operational forecasting of dangerous convective phenomena such as thunderstorm, hail, squall and heavy rainfall is a priority due to the increased frequency of these phenomena and the scale of destruction they bring. However, verification and adjustment of the models are very difficult due to the lack of freely available integrated data on the location and time of the observed phenomena in conjunction with meteorological data used as initial and boundary conditions for numerical models of convective clouds.

At present World Meteorological Organization [2] provides observationally-derived case studies which allow the comparison between the model results and the

B. Murgante et al. (Eds.): ICCSA 2014, Part I, LNCS 8579, pp. 440–451, 2014.

observations of the actual clouds. These data are obtained in the complex field experiments, which are very expensive and extremely rare. The example of such experiment are presented in [3,4]. However it is necessary to carry out a large series of numerical experiments to obtain statistically valid data on the suitability of the model for forecasting. This requires a virtual environment where such experiments could be carried out in an automated mode using a user-friendly interface. Well elaborated database, capable to integrate heterogeneous data from various distributed sources should become the essential component of such environment.

Several firms, research institutions and scientific communities elaborate special systems for meteorological data integration. For example, Raytheon Company [5] has developed Integrated Terminal Weather System (ITWS) that provides automated weather information for use by air traffic controllers and supervisors in airport terminal airspace. ITWS provides the comprehensive current weather situation and highly accurate forecasts of expected weather conditions for next 30 minutes. The ITWS achieves this through integration of data and information from nearly all available sources of meteorological information including NWS (National Weather Service) sensors and weather models, Low Level Wind Shear Alert System (LLWAS), automated weather and surface observing systems and lightning detection systems. Automated weather products produced by the ITWS include wind shear and microburst detection and predictions, storm cell intensity and direction, lightning information and detailed data of the winds in the terminal area.

Advanced Monitoring Methods, LLC [6] utilizes a SQL-based data acquisition system (DAS) for meteorological data collection, validation, graphing, alerts, and reporting. The database structure enables comparative studies and integration of data from different sources and locations.

Institute for Radar Meteorology (IRAM) [7] provides specialized systems for aviation forecasters, collecting data from meteorological satellites, world forecasting centers, meteorological and aviation station networks.

Meteorological Assimilation Data Ingest System (MADIS) [8] is developed by the National Oceanic and Atmospheric Administration (NOAA) Research (Oceanic and Atmospheric Research (OAR)) Earth System Research Laboratory (ESRL) Global Systems Division (GSD) to collect, integrate, quality control (QC), and distribute observations from NOAA and non-NOAA organizations. MADIS ingests data files from NOAA data sources and non-NOAA data providers, decodes the data and then encodes all of the observational data into the common format with uniform observation units and time stamps. Quality control checks are conducted and the integrated data sets are stored in the MADIS database with a series of flags indicating the quality of the observation from the variety of perspectives (e.g. temporal consistency and spatial consistency), or more precisely, the series of flags indicating the results of various QC checks. MADIS users and their applications can then inspect the flags and decide whether or not to use the observation. MADIS data is made available to the enterprise using multiple data transfer protocols via the Internet, including ftp, Unidata's Local Data Manager (LDM) software, or for the surface datasets through the Text/XML Viewer web service found below. Users can subscribe to the entire database, or ask for only particular datasets of interest.

Major part of these data is provided on the commercial basis (ITWS, DAS, IRAM) and thus does not entertain requests for academic research assistance. Besides, the data is stored in the most cases as the static archives on computer disks or other external storage media. And what is the most important matter; all available meteorological data cannot be directly used for verification of convective cloud models due to the absence of integrated information about the place, date and atmospheric conditions during hazardous convective event. We are able to extract the data about atmospheric sounding in a certain place at the certain day, but we are not able to obtain information from the same database or archive about thunderstorm or hail occurrence in that place at that time. We need to look for another information source.

We apply consolidation technology in this paper to create the prototype of system that integrates heterogeneous information about time, location and type of dangerous convective phenomenon in conjunction with the whole set of meteorological data about the state of the atmosphere and the Earth surface. The developed prototype is based on the data freely distributed via the Internet, and thus is available for research organizations which are not able to buy information from commercial sources. The format of the information stored in the database does not require further decoding and can be directly used for simulation of convective cloud characteristics by means of 1.5-D numerical model.

2 Numerical Model of a Convective Cloud

The model is "1.5-dimensional" with the detailed description of microphysical processes. The term "1.5 – dimensional" means the following: though all cloud variables are represented with mean values averaged over the horizontal cross section of the cloud, fluxes in and out of the inner cylinder borders are taken into account.

In the model the region of convective flow is represented by two concentric cylinders [9]. The inner cylinder (with constant radius a) corresponds to the updraft flow region (cloudy region) and the outer cylinder (with constant radius b) – to the surrounding downdraft flow region (cloudless).

Dynamical block of the model is the set of partial differential equations describing evolution in time and space of: vertical component of wind velocity, temperature, mixing ratios of water vapor and cloud droplets under the influence of buoyancy force, gravity, turbulence heat generation/consumption resulted from the phase transitions during the processes of condensation/evaporation.

Microphysical block is represented by the stochastic equation describing the variation of the mass distribution functions of cloud drops, columnar crystals, plate crystals, dendrites, snowflakes, graupel and frozen drops under the influence of the processes of nucleation, condensation, sublimation, coalescence, freezing, melting and breakup both in time and height. The knowledge of distribution function enables to calculate both spectra of hydrometeors and liquid and ice content of a cloud.

Equations are numerically integrated using a finite difference method. Forward-upstream scheme is used. Vertical velocity is averaged over two grid points (point below is taken if $w \geq 0$ or point above if $w < 0$).

Time-splitting method is used for sequential calculations of dynamical and microphysical process. Dynamical processes were calculated at the first stage, microphysical processes at the second one.

Vertical distributions of environmental temperature and relative humidity together with initial impulses of temperature and velocity have been taken as initial conditions. All variables with the exception of temperature and mixing ration of water vapor are equal to zero at the top and at the bottom boundaries of the cylinders.

The model is capable to describe precipitation formation processes in convective clouds under various vertical distributions of temperature and relative humidity of outer atmosphere.

The model can predict maximum and minimum values of dynamical and microphysical characteristics and besides the values of the height of a cloud base and upper boundary, precipitation rate and total quantity of the rainfall. All these characteristics are of major value for prediction of hazardous convective cloud phenomena such as thunderstorms, hails and rain storms.

Detailed description of the model can be found in [10-13].

Qualitative hazard forecasting with the help of the model requires careful verification of the latter on real data. It is necessary to check whether the parameters calculated by the model fit the conditions of the thunderstorms. We have to take real vertical distributions of the temperature and humidity in those days, when thunderstorms have been registered, provide the model calculations using the obtained data as initial conditions and to identify whether the calculated parameters correspond to the observed thundercloud parameters.

A large number of model runs is necessary to obtain statistically justified data about model suitability for forecasting. The task is very time-consuming if it is implemented manually. The software environment has to be developed where such experiments could be carried out in automated mode using a user-friendly interface.

The database containing all necessary meteorological information should be one of the elements of such environment. The database is to consolidate information from different sources and store it in a proper format that can be used as the boundary and initial conditions for the model runs.

We have elaborated relational database which scheme is presented in the form of tables and relationships between them. The "tables" are the realization of the entities and fields in the tables are the realization of the properties of the entities. It should be noted that the developed logical model of the database is in the third normal form because of the absence of transitive dependencies.

The elaborated data scheme contains 9 "tables" and can be realized using SQL. During the physical database modeling we describe the data types for each type of stored information as well as the ways and the place of their physical location. Moreover, for each field in the table data types have been defined which are the most suitable for storing of the relevant information. Final Entity-Relationship (ER) diagram is shown in Fig. 1

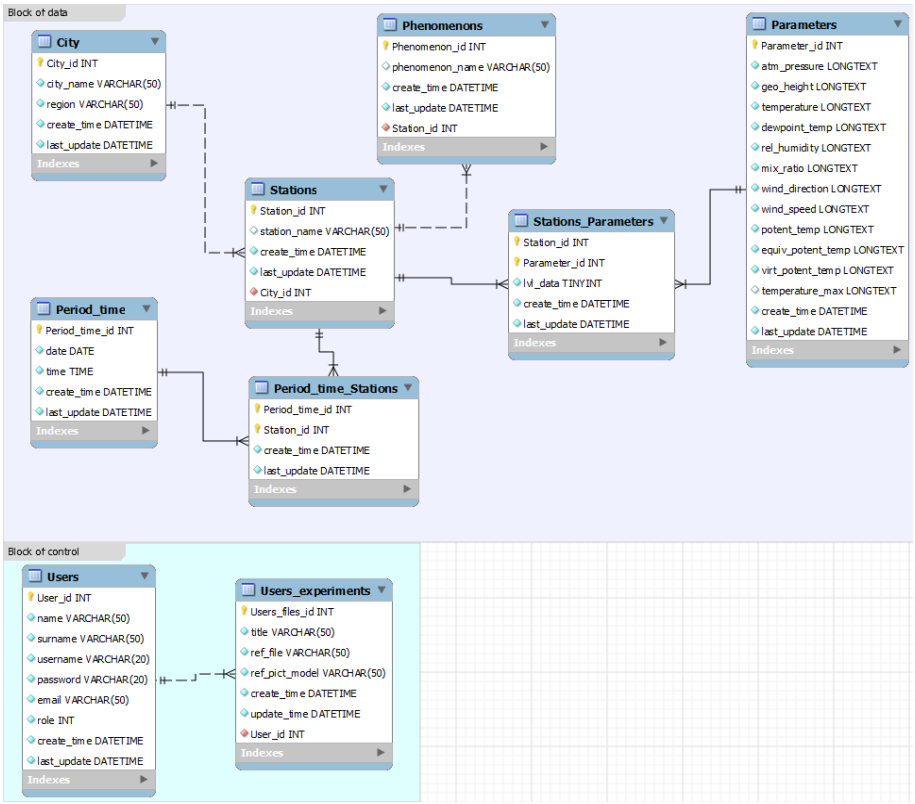

Fig. 1. ER diagram

3 System for Heterogeneous Data Selection and Transformation

As it has been mentioned above meteorological information needed for model verification can be found in different heterogeneous sources where it is stored in different formats. We provide the selection of the sources using the following principals: the sources should be free; they should contain information about the place, the date and the type of hazardous event and information about the vertical distribution of the temperature and humidity in the environment atmosphere on the date and in the place of the event. Additional information about the pressure and the maximum temperature on the surface, type of clouds and synoptic situation is desirable, but not required. We have chosen two of all the diverse sources of meteorological data which meet the above criteria.

Internet sites http://weather.uwyo.edu/ and http://meteocenter.net/ together allow to obtain the whole amount of the relevant information. The site of "Meteocenter" http://meteocenter.net/ provides the archive data which can be used for extracting information about the type of hazardous event, the place and the date where it has occurred. The site of the University of Wyoming (http://weather.uwyo.edu/) allows

to obtain vertical distributions of atmospheric temperature and humidity. So using the information from the first site as the input data, one can extract the required data from the second site.

Sites vary in their structure, user interface languages, structure and content of provided meteorological information. Therefore it is not possible to create the universal software running flawlessly for both sites. We have to develop special application (the system) capable to extract the necessary data from heterogeneous sources. We use the following tools for the system development: Php 5.5, MySQL 5.1 as the backend and Html 5, Css 3, Javascript (ECMAScript 5) as the client side.

We have selected server PHP programming language version 5.5, which has the ability to use the libcurl library. This library allows interaction with many different servers on multiple protocols, speeding up the data retrieval from the sites significantly due to multithreading and multitasking.

Our system contains several sophisticated components providing heterogeneous data transformation into the single structure. The whole algorithm of the system is presented in Fig. 2.

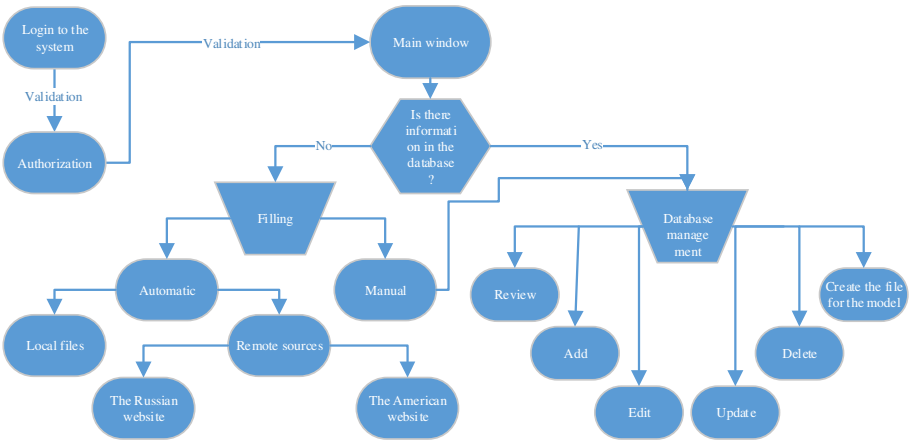

Fig. 2. The algorithm of system operation

After authorization an expert starts working with the system either by selecting new data or by using the information which has already been selected in the previous sessions. It is possible to choose the place (the city), the date (year, month, date) and the hazardous event of interest.

The second step concerns rapid information receipt with the help of multithreading technique. The process is fully automated. Three main stages i.e. extraction, transformation and loading (ETL) of the data occur invisibly to the user.

After extraction, the data falls into a temporary file for a short time, where it is converted into the required format, and then inserted into the database, operated by an expert with the help of web application. Let us consider the ETL stages noted above.

4 ETL Stages Description

After successful login the user has to choose: either to insert the necessary information into the database, which can be done manually or automatically, or to start managing the data stored in the database. Let us consider the case of auto-fill database by extracting information from the two sites mentioned above.

The selected sites have their own systems of data uploading. These systems allow us the access to all the information stored on the sites in the public domain. On the one hand, it simplifies the programming part, because we do not need to develop our own data upload system, but on the other hand, we have no choice and we can only adjust the form proposed by the sites.

Extraction starts with the Russian site, as some of the information obtained from it serves as an input to the U.S. site. The user inserts relevant parameters such as time, date, city and the type of hazardous event. It is not possible to get data from the Russian site directly, so we use the library *libcurl*, which will send the required query.

When the data is extracted, we obtain the file in *CSV* (Comma-SeparatedValues) format, allowed for unloading from Russian site. Information is saved in a temporary file and the session is closed with the release of resources.

To extract data from the U.S. site, we need to aggregate information from Russian site using the parameter *time*. The *time* is due to the presence of the parameter *event*. We use an array obtained in aggregation step and select the time intervals interesting for the user. Then we copy them into an associative array, that contains information about the days and the corresponding periods, and create an array of references-queries with the parameters denoting the type, the place and time of the event. Then we test the query by splitting it into the components by using function *parse_url($ref_req[$i], PHP_URL_QUERY*). We check each component on the consistency, absence of spaces and special symbols.

Parallel data extraction from the site is implemented by creating the array of descriptors and the array of query results. At the end of extraction, we integrate the results and release the memory from multithread mechanism. Final structured data is stored in temporary txt file $out_txt, where each line is strictly of 77 characters.

Data transformation is the next step after the extraction. We need to prepare the data for loading into the database and converting it to the form which is most convenient for analysis. The conversion process consists of five steps. Let us consider them in more detail.

Since the data may vary in their structure and format, we need to restructure it that is to bring all the data into a form relevant for our analysis. In our case, the data from the different sites can contain special characters and omissions which should be abandoned or converted to the special symbols capable to be recognized by the code of the numerical model. Since omissions which may critically affect the forecast of convective clouds are contained only at the U.S. site, we apply the following: find all the lines in the file obtained after the 1st stage, with the function *preg_match_all ('/ (. *) {77} / ', $ matches [1] [0], $ matches_count_str)*, which is fed to the input of the search pattern, the place of event and a variable that will store the result. We replace the blank cells by the desired values and then check numerical data and get it using function *preg_match_all*, which forms the array *$matches_values*, containing the parameters to be used in further processing.

Processing the data of the Russian site is much easier. We simply apply parser for CSV format and get the data in the array *$values*. Then we select the first, the second, the sixth and the seventeenth columns of the array, which are responsible for the time period, the date, the event and the maximum temperature respectively. We additionally create an associative array of hazardous events and use the array to convert the number of the event in its name and vice versa.

The last step of ETL - data loading into the database is implemented after the completion of the processes of extraction, transformation and cleansing. Loading process comprises data transition from the staging area (files) into the database structure . Since all the parameters are already stored in the array *$merge_data*, we load them into the database using sql- queries. Loading is also possible from the local sources with the specific file structure, such as CSV, as in the Russian site and TXT, as in the U.S. site. The scheme of ETL stages is presented in Fig. 2.

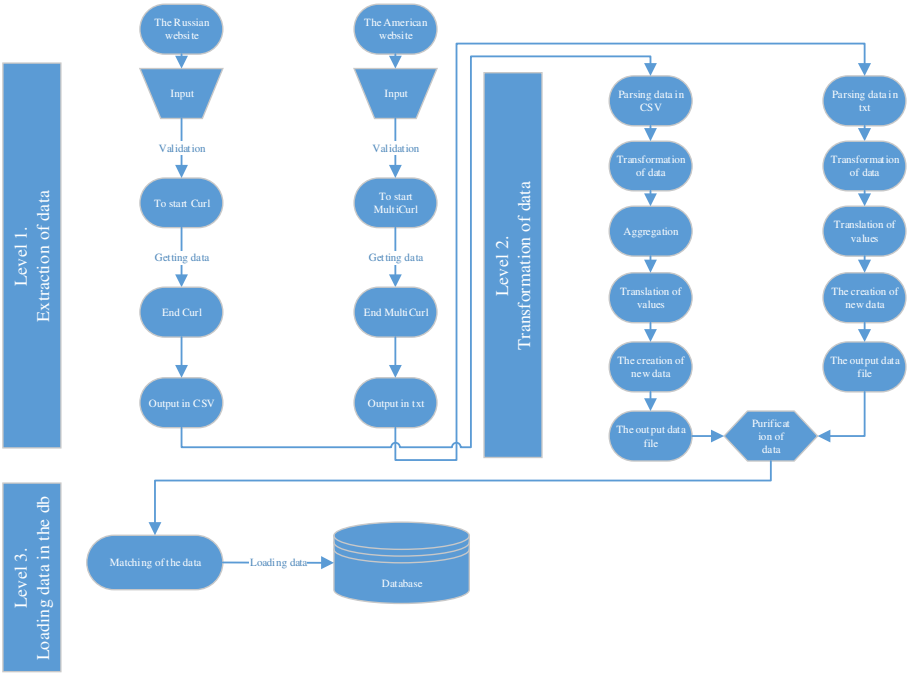

Fig. 3. ETL stages description

After loading, which has been carried out in automatic mode, database control is transferred to an expert, who can provide a full set of data operations: view, add, edit, update, delete and create a special input file for numerical model of the convective cloud.

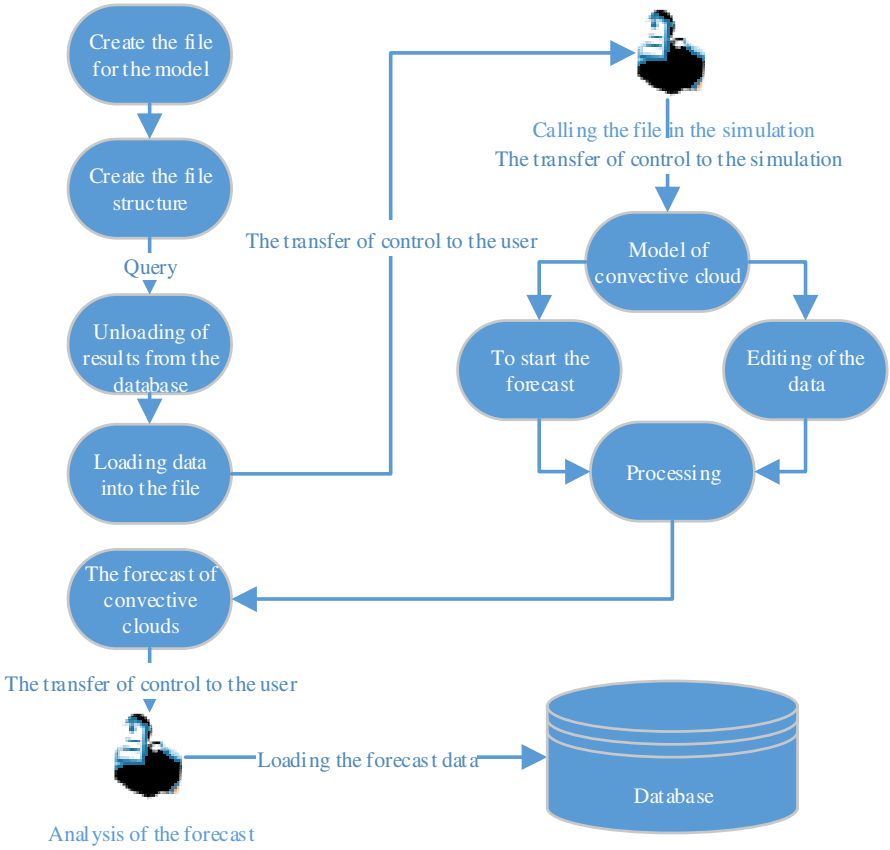

Fig. 4. Using the database for forecasting

After creating the file the expert can run exe file of the convective cloud numerical model and analyze the results of model simulation. Initial input file can be changed by the model program. Modified file can be either loaded to the database or saved on the expert computer. The scheme of the described process is presented in Fig.3.

Database has been created specifically for the application and meets all the necessary criteria. It allows you to search for specific parameters, such as hazardous event, thus freeing us from the manual search. The database is connected to the web- application which allows analytical work with the data.

Two screenshots illustrating the process of the whole system operation are presented in Fig. 4,5.

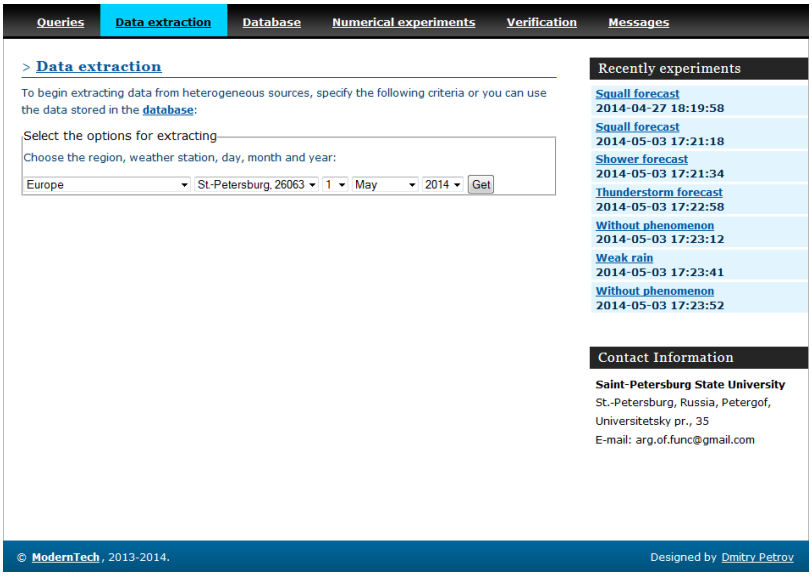

Fig. 5. Screenshot illustrating the process of data extraction

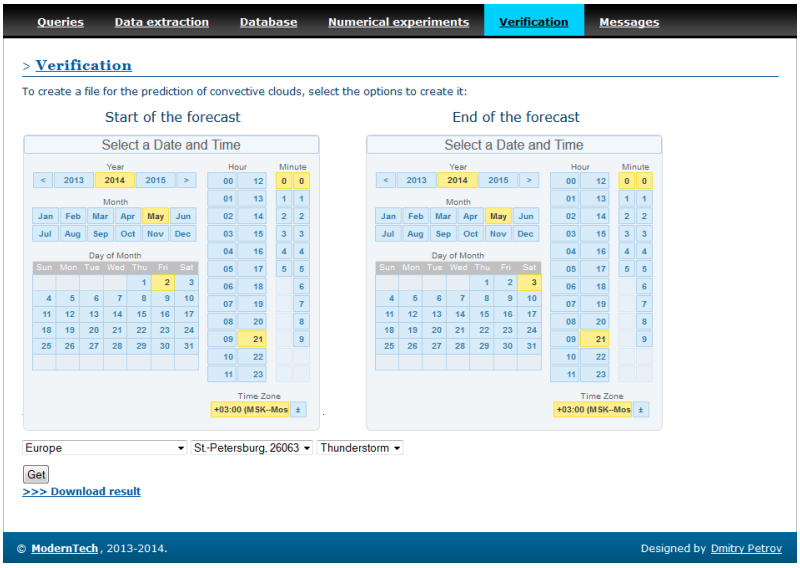

Fig. 6. The screenshot illustrating the start of model verification process

5 Conclusions

Special application (the system) capable to extract the necessary data from heterogeneous sources of meteorological information has been developed. The system contains several sophisticated components providing heterogeneous data transformation into the single structure.

Three main stages of data consolidation extraction, transformation and loading (ETL) are provided automatically and invisibly for the user. The whole algorithm of the stage realization is described in detail.

After extraction, the data falls into a temporary file for a short time, where it is converted into the required format, and then inserted into the relational database, operated by an expert with the help of web application.

The database contains the whole set of meteorological information about the state of the atmosphere at the place and at the time when a hazardous convective phenomenon is recorded.

System development has been provided with the help of Php 5.5, MySQL 5.1 serving as the backend and Html 5, Css 3, Javascript (ECMAScript 5) serving as the client side.

Rapid extraction of the relevant data has been provided by using multithreading technique supported by the functions of the libcurl library.

The developed prototype is based on the data freely distributed via the Internet, and thus available for research organizations which are not able to buy information from commercial sources.

The format of the information stored in the database does not require further decoding and can be directly used for simulation of convective cloud characteristics by means of 1.5-D numerical model.

We plan to use elaborated system for validation of 2-D [14] and 3-D [15] models in future, by adding new information parameters into the database, referring to space distributions of radar reflectivity and wind velocity.

The developed system can be used for providing a large number of numerical experiments using the real atmospheric parameters as the input data. The results of the experiments allow to obtain statistically justified data about efficiency of cloud model usage for forecasting hazardous convective events such as thunderstorms, heavy rains and hail. Such model validation could be carried out in of automated mode using a user-friendly interface.

Acknowledgment. This research was sponsored by the Saint-Petersburg State University under the project 0.37.155.2014 “Research in the field of designing and implementing effective computational simulation for hydrophisical and hydro-meteorological processes of Baltic Sea (and the open Ocean and offshores of Russia)”

References

1. Seifert, A., Baldauf, M., Stephan, K., Blahak, U., Beheng, K.: The Challenge of Convective-Scale Quantitative Precipitation Forecasting. In: Proceedings of the 15th International Conference on Clouds and Precipitation, ICCP 2008 (2008)
2. 8th International Cloud Modeling Workshop, http://www.atmos.washington.edu/~andreasm/workshop2012/cases.html
3. Heldson Jr., J.H., Farley, R.D.: A numerical modelling study of a Montana thunderstorm. Part I: Model results versus observations. JGR D5, 5645–5660 (1987)
4. Dye, J.E., Jones, J.J., Winn, W.P., Cerni, T.A., Gardiner, B., Lamb, D., Pitter, R.L., Hallett, J., Saunders, C.P.R.: Early Electrification and Precipitation Development in a Small, Isolated Montana Cumulonimbus. J. of Geophys. Res. 91(D1), 1231–1247 (1986)
5. Integrated Terminal Weather System (ITWS), http://www.raytheon.com/capabilities/products/itws/
6. Advanced Monitoring Methods. High Quality Data Collection from any Location, http://advm2.com/
7. Institute for Radar Meteorology, http://www.iram.ru/iram/index_en.php
8. Meteorological Assimilation Data Ingest System (MADIS), http://madis.noaa.gov/
9. Asai, T., Kasahara, A.: A Theoretical Study of the Compensating Downward Motions Associated with Cumulus Clouds. Journal of the Atmospheric Sciences 24, 487–497 (1967)
10. Raba, N.O., Stankova, E.N.: Research of influence of compensating descending flow on cloud's life cycle by means of 1.5-dimensional model with 2 cylinders. In: Proceedings of MGO, vol. 559, pp. 192–209 (2009) (in Russian)
11. Raba, N., Stankova, E.: On the Possibilities of Multi-core Processor Use for Real-Time Forecast of Dangerous Convective Phenomena. In: Taniar, D., Gervasi, O., Murgante, B., Pardede, E., Apduhan, B.O. (eds.) ICCSA 2010, Part II. LNCS, vol. 6017, pp. 130–138. Springer, Heidelberg (2010)
12. Raba, N.O., Stankova, E.N.: On the Problem of Numerical Modeling of Dangerous Convective Phenomena: Possibilities of Real-Time Forecast with the Help of Multi-core Processors. In: Murgante, B., Gervasi, O., Iglesias, A., Taniar, D., Apduhan, B.O. (eds.) ICCSA 2011, Part V. LNCS, vol. 6786, pp. 633–642. Springer, Heidelberg (2011)
13. Raba, N.O., Stankova, E.N.: On the Effectiveness of Using the GPU for Numerical Solution of Stochastic Collection Equation. In: Murgante, B., Misra, S., Carlini, M., Torre, C.M., Nguyen, H.-Q., Taniar, D., Apduhan, B.O., Gervasi, O. (eds.) ICCSA 2013, Part V. LNCS, vol. 7975, pp. 248–258. Springer, Heidelberg (2013)
14. Khain, A., Pokrovsky, A., Pinsky, M.: Simulation of Effects of Atmospheric Aerosols on Deep Turbulent Convective Clouds Using a Spectral Microphysics Mixed-Phase Cumulus Cloud Model. Part I: Model Description and Possible Applications. J. Atm. Sci. 61, 2963–2982 (2004)
15. Morozov, V.N., Veremey, N.E., Dovgalyuk, Y.A.: Modeling of the Electrification of 3d Convective Clouds (Review). In: Proceedings of Voeikov Main Geophysical Observatory, vol. 559, pp. 231–359 (2009)