Global dataset for the parameterisation of lakes in Numerical Weather Prediction and climate modelling

E. Kourzeneva, CRNM/GAME, Météo-France CNRS, Toulouse, France,
and RSHU, St. Petersbourg, Russia

1. Introduction

The structure of the atmospheric boundary layer depends on physical state of different types of the underlying surface, including lakes. In regions with high percentage of lake area, lakes affect local weather conditions and a regional climate. The problem becomes particularly pressing as the horizontal resolution of atmospheric models increases. Lakes should be parameterized in Numerical Weather Prediction (NWP) and climate models, and for this we need fields of external lake parameters. Fields should be global and, in principle, should contain information about properties of all existing lakes. Any atmospheric model with a lake parameterization scheme included needs at least information about lake depth, the mean lake depth or even the bathymetry. Great fidelity of the depth data is not critical, but global coverage is important. The lake fraction (the percentage of the atmospheric model grid box covered by lake water) is another external parameter needed by an atmospheric model.

Different lake databases are developed for different purposes. Regional databases are concentrated on individual characteristics of lakes but they do not represent all information on one map. Global databases pay much attention to the detailed information about geographical location of lakes, their extent and distribution but without providing individual physical characteristics (except from very large lakes or lakes significant from socio-economic point of view). See, for example, the Global Lakes and Wetlands Database (GLWD) (Lehner and Döll, 2004), the Global Land Cover Characteristics dataset (GLCC) (Loveland et al., 2000), the ECOCLIMAP dataset (Masson et al., 2003). Being represented in the raster form with pixels classified as "inland water"/"no inland water", these databases can be used as a map. The lake fraction can be calculated from such a map in a standard way.

The dataset presented in this paper provides the external parameters fields for the parameterisation of lakes in atmospheric modelling. It combines depth information for the individual lakes from different sources with a map. As a result, lake depth is represented on the global grid with the resolution of 30 sec. of arc (approx. 1 km). For some large lakes the bathymetry is included. Additionally, the software to project the lake-related information accurately onto an atmospheric model grid is provided. The prototype for this dataset was developed for Europe and is described in details in (Kourzeneva, 2010).

2. Data sources

2.1. Mean depth information for individual lakes

Data for individual lakes were collected from different regional databases and water cadastres. For Europe, different organizations kindly provided data, mainly through personal communication, see (Kourzeneva, 2010) for details. For the rest of the world data were extracted from different sources in internet. Often we relied on data from Wikipedia, mainly from its national pages, which for some countries are very rich. Although Wikipedia is the "semi-scientific" source of information and provides no legal warranty, we did not reject this data. The reason is that for Wikipedia people use information from many scientific and governmental

institutions around the world and most of pages contain references to the appropriate publications, but it is difficult to contact these organizations directly.

Both natural and manmade lakes are considered. Special attention is paid to saline lakes and endorheic basins. Freshwater lake models can't describe their behavior. They can change size and shape over time. Some of them are intermittent or ephemeral. Saline lakes are separated from freshwater ones and form the additional dataset. Lakes with low salinity (less then 10 ‰) and with stable size and shape are considered as freshwater. By now, the main dataset comprises about 13 000 freshwater lakes, the additional dataset comprises about 220 saline lakes and endothreic basins. The list of data sources includes ca. 295 references, and they are located together with data.

For each individual lake we used the following information: geographical coordinates of a point on the water surface, the mean depth of the lake, its maximum depth, its surface area, the lake name and the name of the country where the lake is located. Where the data about the mean lake depth were missing, the default value of 10 m was used.

2.2. Map for lake depth information

At present, in geophysical sciences much attention in paid to the development of global and regional ecosystem datasets - GLCC, ECOCLIMAP, GLC2000 (Bertholomeé and Belward, 2005), CORINE (CEC, 1993), GLOBCOVER (Bicheron, 2006). They are used by atmospheric models to specify fields of external parameters. They have different resolution (25 m – 1 km) and some of them distinguish between different types of water bodies – seas, lakes, rivers. However, as it was discussed in literature (Lehner and Döll, 2004; Merchant and MacCallum, 2009; Kourzeneva, 2010), most of them have inaccuracies in the shoreline. These inaccuracies are inherited from the initial data sources, as most of them use the Digital Chart of the World (DCW), (ESRI, 1993) and the ArcWorld 1:3M dataset (ESRI, 1992) to specify the shoreline. To get rid of these inaccuracies, the high resolution remote sensing could be helpful, but the correct automatic classification based on space-born data only is difficult. In order to choose the ecosystem dataset for a basic map, we made the express-comparison of 4 global products with 1 km resolution. These are GLCC, GLWD, ECOCLIMAP and ECOCLIMAP2 (Faroux et al., 2009; Champeaux et al., 2004). The comparison was based on visual estimates. The remote sensing data were used as a gage. We examined several test regions on the globe with the main attention given to Europe. Artifacts (e.g. a big lake does not exist on the map but do exist in reality, or there is a false big lake or a false island on the map) and the bias (too much water/too few water) were estimated. After removing some artifacts from ECOCLIMAP2, it was chosen for mapping of lake depth information. ECOCLIMAP2 distinguishes between rivers and lakes, but many rivers are erroneously referred as chains of lakes.

2.3. Bathymetry data for large lakes

At present, there are two global datasets containing the bathymetry information for large lakes. The dataset ETOPO1 (Amante and Eakins, 2009) has the resolution of 1 min of arc and contains the detailed information about the bathymetry of Great Lakes. This information was used. The dataset ETOPO5 (ETOPO5, 1988) has the resolution of 5 min of arc and contains the bathymetry information also for some other large lakes apart from Great Lakes. But the quality of data is quite poor, so we refrained from using this dataset. The bathymetry for 30 other large lakes (apart from Great Lakes) was obtained from topographic and navigation maps in a graphic form by digitizing with kriging interpolation method used for gridding. Topographic and navigation maps were obtained from different sources, many sketch-maps were taken from the International Lake Environmental Committee database (ILEC, 1988-1993). Note that the model variable

which communicates information between the lake and the atmosphere is the lake surface temperature. Its sensitivity to the lake depth is quite low for very deep lakes. In the lake model FLake (Mironov, 2008) which is used in many NWP and climate models to parameterize lakes, there is a limit to the lake depth of 50 m. So, the bathymetry was included for large lakes which are not too deep (the mean depth is less then 70 m), not too shallow (the maximum depth is more then 10 m), and have the difference between the mean depth and the maximum depth of more then 6 m. In the other words, the bathymetry is not included for the lakes which can be in practice characterized by their mean depth. So, the bathymetry for such big lakes as Lake Baikal, Lake Tanganyika, Lake Chad, Lake Balaton, and Lake Manitoba is not included.

3. Methodology

The methodology to combine automatically mean depth data for individual lakes with a raster map described in (Kourzeneva, 2010) was further developed and used. Its basic ideas are:
-       A lake on a raster map (a "spot-lake") is a set of conterminal pixels with the "lake" ecosystem type. Our task is to find correspondences between "spot-lakes" and lakes in the dataset for individual lakes.
-       The dataset for individual lakes may have random errors in coordinates of a point on the lake water surface; the shoreline on a map is also defined with random errors. So, the probabilistic approach was used.

The new algorithm is described here briefly.
-    For the lake H from the dataset for individual lakes we considered the coordinate vector X of a point on its surface as a continuous random value with the normal distribution. We assume that in the dataset for individual lakes its mean value X_0 is given. We prescribe the value of variance and calculate the field of probability P_h of the hit of this point into every pixel of the raster map within some influence radius around X_0.
-    We assume that the "spot-lake" L on the raster map corresponds to a lake L in reality. In the pixels of the raster map we appoint the field of probability P_b of the event that the pixel in question belongs to the lake L in reality. The field is constructed so that P_b decreases according to the square-law in the vicinities of the shoreline of the "spot-lake" L.
-    For every lake H from the dataset for individual lakes we find the pixel on the raster map corresponding to X_0. In the area around this pixel we calculate the probability field P_h. For every "spot-lake" L on the raster map within this area we calculate also the probability field P_b. The total probability P that the lake H is the same lake with the "spot-lake" L is $P = P_h \cdot P_b$. We find the maximum field value of P and set the correspondence between the lake H and the "spot-lake" L having the probability P.
-    As a result of the previous step, every "spot-lake" L on the global raster map receives more than one correspondence with a lake H or it does not receive any correspondence. In the case of zero correspondence (the "spot-lake" L was not recognized), every pixel of the "spot-lake" L receives the default depth value. In the case of more than one correspondence we choose that with the maximum probability P value, and every pixel of the "spot-lake" L receives the depth value from the appropriate lake H.

We used the 15 km value for the influence radius and the 10 m value for the default lake depth. The same default depth value of 10 m was used for the lakes with missing lake depth information in the dataset for individual lakes. All pixels of the raster map with the "river" ecosystem type received the default depth value of 3 m. At the moment, only information from the dataset for freshwater lakes was used, the saline lakes were not included.

We applied the mapping method twice. First, preliminary run made possible to find and to fix the rough errors in the coordinates of large lakes in the dataset for individual lakes. The final product was obtained after the second run.

The bathymetry for large lakes was first interpolated into the grid of our raster map with the 30 sec. of arc resolution. The simple linear interpolation was used. Then the shoreline for every large lake was put into accordance with our raster map ECOCLIMAP2. The nearest-neighbor method was used for extrapolation if necessary. Finally, for the large lakes we replaced the mean depth lake values in every pixel by the bathymetry.

4. Products

1) The global gridded dataset containing lake depth information, namely the mean lake depth values or the bathymetry, with the resolution of 30 sec. of arc (approx. 1 km).
2) The additional dataset containing the variable S to estimate reliability of the lake depth information in every pixel of the grid. This variable is determined as follows. $S=0$ if there is no inland water, $S=1$ if the "spot-lake" was not recognized, $S=2$ if the "spot-lake" was recognized but with missing lake depth information in the dataset for individual lakes, $S=3$ if the real depth value was used, $S=4$ if there is a river and the default depth value for rivers was used. This dataset can be useful if we want to estimate the quality of our data.

These products are possible to download freely from the lake model FLake web page (*http://nwpi.krc.karelia.ru/flake/*). The datasets for individual lakes, freshwater and saline, as well as the list of lakes with the included bathymetry, all provided with references to the data sources, can be also downloaded from this web page. Illustrations for the gridded lake depth dataset are presented in Figs. 1-2 with the visualized lake depth data for the areas near Great Lakes and in Sweden.

5. Projection onto an atmospheric model grid

The lake depth field is discontinuous hence averaging of the lake depth values is incorrect. The method to aggregate the lake depth information onto an atmospheric model grid, which is in principle coarser than the grid for lake depth, was described in (Kourzeneva, 2010). The method is based on the empirical probability density functions for every grid box and uses the mode statistics (the most probable lake depth value for the grid box in question). This method is recommended also to apply for the presented gridded lake depth field with the fine resolution of 30 sec. of arc to project it onto the atmospheric model grid.

The appropriate software was developed and also can be downloaded from (*http://nwpi.krc.karelia.ru/flake/*). Different atmospheric models use very different coordinate systems, map projections and have very different grids. Hence, it is very difficult (if possible at all) to have the universal software, which does not need any additional efforts in programming from a user. So, the FORTRAN90 routine is provided to aggregate the lake depth data for one grid box of the atmospheric model (target) grid approximated by the polygon in geographical (longitude and latitude) coordinates. The output from this routine is the lake fraction for the grid box in question, the most probable depth of lakes in the grid box in question, and the most probable value of the variable S (see above) for the grid box in question. If somebody prefers to use the average value instead of the most probable, this option is also possible. The examples of the output from this software, namely the fields of lake fraction and lake depth are shown in

Figs. 3-4. For the target grid the rotated spherical coordinates are used with the new South Pole location in the point with geographical coordinates of 30° in longitude and of -30° in latitude, the resolution is 0.1°. The domain covers the area around Baltic Sea. It includes large lakes Lake Ladoga, Lake Onega, Lake Vanern, Lake Vattern, Rybinskoe Reservoir, and Lake Peipsi (Chudskoe).

6. Discussion

The automatic mapping method makes it very easy to include new lake data and to update the product. The quality of the final product is strongly dependent on presence of data in the dataset for individual lakes. It is very important to maintain it in future, adding new lakes and correcting the data.

Saline lakes should be also taken into account, bearing in mind their specific features. Rivers are defined in ecosystem datasets very poorly (except from GLWD, the situation is better there). Sometimes it is very difficult to distinguish automatically on the map the boundary between the river and the lake or between the river and the sea. Coastal lagoons, even freshwater, are treated by ecosystem datasets as "sea water" very often. In many cases distinguishing between different types of water bodies is difficult, as the definition of lake in reality is rather questionable (Lehner and Döll, 2004; Merchant and MacCallum, 2009).

Only express-comparison of the different raster maps was made. Better comparison would be useful. New raster maps will appear in future with the shoreline described more precisely (Bicheron et al., 2006; Merchant and MacCallum, 2009). They also could be used for mapping. The automatic method of mapping makes it possible to change easily the ecosystem dataset used for a raster map.

Note that the accuracy of the bathymetry data in the presented product is low and suitable only for atmospheric modeling, hydrological or environmental applications, but not for navigation. New bathymetry information for large lakes can be easily included, if we have the appropriate data. Bathymetry maps for large lakes in digital or graphic form do exist, although many of them are not free. This information should be included also.

Even if we could collect all the measured data on lake depth, this is not enough. In some regions (e.g. Northern Canada, Siberia) the depth was not measured at all for many lakes. So, indirect estimates, e.g. from the orography variation or from the surface temperature annual cycle, are very welcome. As least, the default lake depth value may depend on a region.

7. Conclusion

The new Global dataset for the parameterisation of lakes in Numerical Weather Prediction and climate modelling is presented. It contains global gridded data for lake depth, the mean values or the bathymetry, with the resolution of 30 sec. of arc and the additional dataset about the reliability of the depth data. They were obtained by mapping the information from the dataset for individual lakes comprising ca. 13 000 lakes, to the map of dataset for ecosystems ECOCLIMAP2 (Faroux et al., 2009; Champeaux et al., 2004). For mapping, the new method of appointed probabilities was used. The method is automatic, it allows easy maintenance of the product and provides good tools for further developments. The new lake depth data are highly desirable. To project the presented gridded lake depth data onto an atmospheric model grid, the

method of empirical probability density functions is recommended. The appropriate software (FORTRAN90 routine) is provided.

References:

Amante, C. and B.W. Eakins, 2009: ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. *NOAA Technical Memorandum NESDIS NGDC-24*, 2009-3, 1-19 (data available online at http://www.ngdc.noaa.gov/mgg/global/global.html).

Bertholomeé, E. and A.S. Belward, 2005: GLC2000: a new approach to global land cover mapping from Earth observation data. *Int. I. Remote Sensing* 26(9/10), 1959-1977.

Bicheron, P. et al., 2006: GLOBCOVER: a 300m global land cover product for 2005 using ENVISAT/MERIS time series. *Proceedings of the Recent Advances in Quantitative Remote Sensing Symposium*, Valencia, September 2006.

CEC, 1993: CORINE Land Cover technical guide. European Union. Directorate-Generale Environment, Nuclear Safety and Civil Protection. Luxembourg.

Champeaux, J.-L., K.-S. Han, D. Arcos, F. Habets and V. Masson, 2004: Ecoclimap2: A new Approach at Global and European Scale for Ecosystems Mapping and Associated Surface Parameters Database Using SPOT/VEGETATION Data - First Results. *International Geosciences and Remote Sensing Symposium*, 3, 2046-2049.

ETOPO5: Data Announcement 88-MGG-02, Digital relief of the Surface of the Earth. NOAA, National Geophysical Data Center, Boulder, Colorado, 1988. (also available online at http://www.ngdc.noaa.gov/mgg/global/etopo5.html).

ESRI: Environmental Systems Research Institute, 1993. Digital Chart of the World 1:1 Mio. Redlands, CA. Data obtained on 4 CDs (also available online at http://www.maproom.psu.edu/dcw/).

ESRI: Environmental Systems Research Institute, 1992. ArcWorld 1:3 Mio. Continental Coverage. Redlands, CA. Data obtained on CD.

Faroux, S., J.-L. Roujean, A. Kaptue, V. Masson, 2009: La base de donnees de parameters de surface ECOCLIMAP-II sur l'Europe. *Note de travail du groupe de meteorology a moyenne echelle*. N 86. Meteo-France / C.N.R.M. / P. A. D., 1-120.

ILEC: International Lake Environmental Committee, 1988-1993. Survey of the State of World Lakes. *Data books of the world lake environments*, vols. 1-5. ILEC/UNEP Publications, Otsu, Japan (also available online at http://wldb.ilec.or.jp/).

Kourzeneva, E., 2010: External data for lake parameterization in Numerical Weather Prediction and climate modeling. *Boreal Env. Res.* 15, 165–177.

Lehner, B. and P. Döll, 2004: Development and validation of a global database of lakes, reservoirs and wetlands. *Journal of Hydrology* 296/1-4, 1-22.

Loveland, T.R., B.C. Reed, J.F. Brown, D.O. Ohlen, J. Zhu, L. Yang and J.W. Merchant, 2000: Development of a global land cover characteristics database and IGBP DISCover from 1-km AVHRR data. *Int. I. Remote Sensing* 21(6/7), 1303-1330. (data and documentation available online at http://edc2.usgs.gov/glcc/glcc.php).

Masson V., J.L. Clampeaux, F. Chauvin, C. Meriguet and R. Lacaze, 2003: A global database of land surface parameters at 1-km resolution in meteorlogical and climate models. *J. Climate* 16, 1261-1282.

Merchant C. J. and S.N. MacCallum, 2009: *ARCLake Quaterly Report no. 1.* The University of Edinburg: 1-13. (data available at http://www.geos.ed.ac.uk/arclake).

Mironov D., 2008: Parameterization of lakes in numerical weather prediction. Description of a lake model. *COSMO Technical Report* 11. Deutscher Wetterdienst, Offenbach am Main, Germany, 1-41.
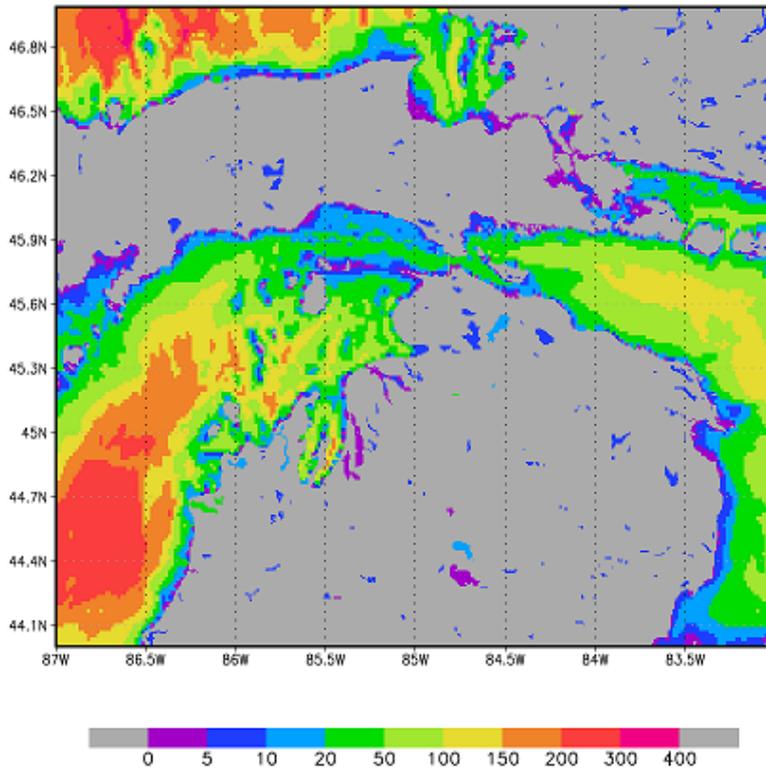
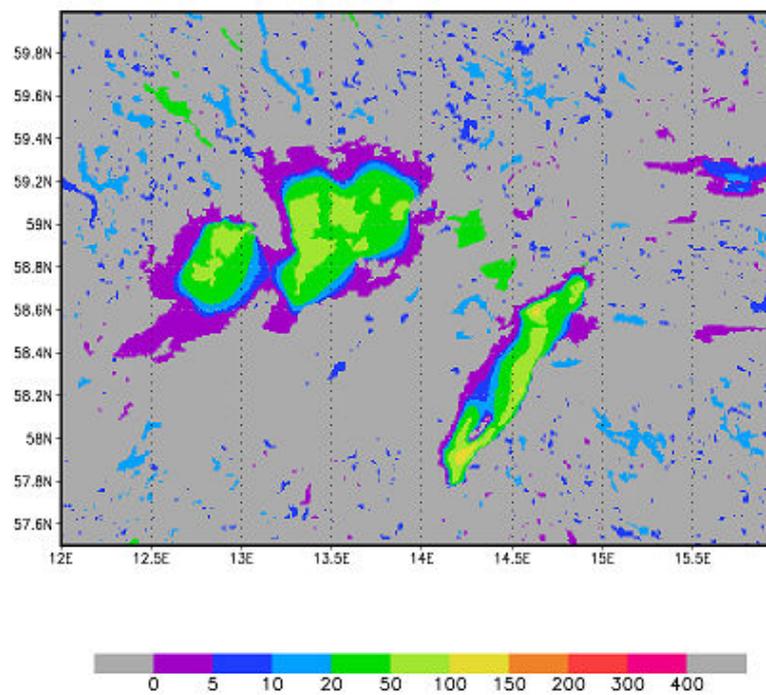Fig. 1. Lake depth, m for the area near Great Lakes on the grid with 30 sec. of arc resolution

Fig. 2. The same with Fig. 1 but for the area in Sweden including Lake Vanern and Lake Vattern
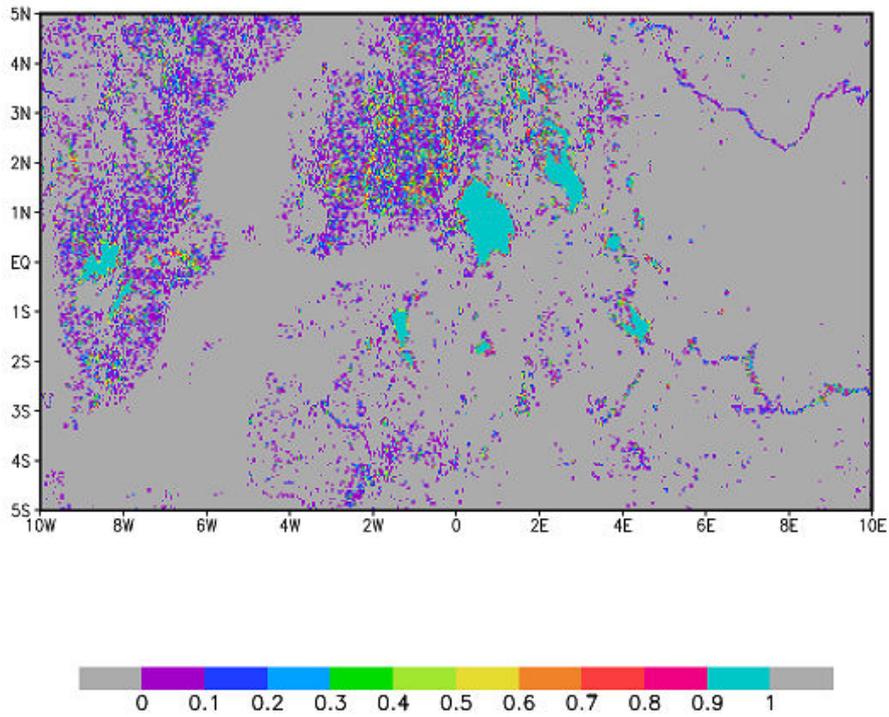
Fig. 3. Lake fraction (0-1) for the atmospheric model grid with the resolution of 0.1° (see text for details), the domain covers the area around Baltic Sea
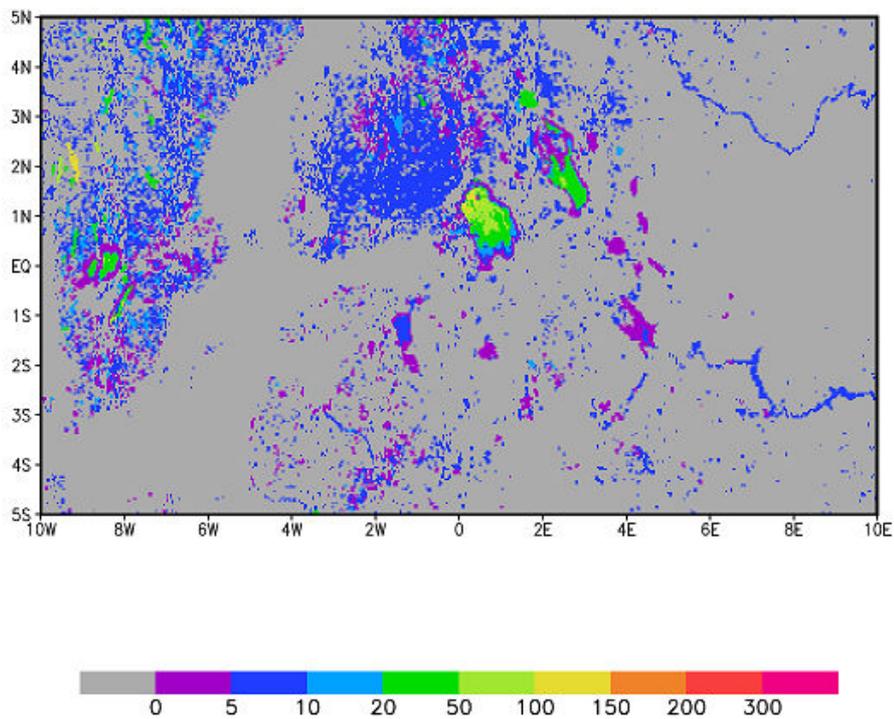

Fig. 4. Mean lake depth in meters for the atmospheric model grid with the resolution of 0.1° (see text for details), the domain covers the area around Baltic Sea