Jaroslav CHYBA[1] **M 5**

PROBLEMS OF PREDICTION IN EARTH SCIENCES

[1] JUDr. RNDr. Jaroslav Chyba, DrSc., Povodí Moravy, s.p, . Dřevařská 11, 601 75 Brno, Czech Republic chyba@povodi.cz

1. Introduction

Prediction has been the highest aim of human knowledge since the beginning of ancient civilizations. Prediction of natural phenomena like position of planets, weather or floods was the basis of the power of priests in ancient Egypt and Mesopotamia.

Recent methods of prediction are based on mathematical description and physical explanation of experiments. Modern methods have started with the work of Galileo Galilei, Isaac Newton, Johannes Kepler and other founding fathers of science.

Methods of prediction in Earth Sciences are heterogeneous; however, some common features may be outlined.

I was busy with prediction of natural phenomena as prospecting geophysicist for one half of my professional life and as hydrologist for the other half. In the following text, I will present a review of mathematical tools and applications. Subsequently, I will point out several mathematical, technical and legal problems connected with prediction.

1. Review of methods

2. 1. Prediction in steady-state systems

Prediction in steady-state systems relates to the development of human knowledge, not to the development of the system itself. Example: prediction of the occurrence of mineral deposits in a given area.

1.2. Prediction in time-variable systems

Prediction in time-variable systems can be based on physical concepts or on a statistical analysis of past data. This bipartition is useful for a rough classification of methods, but it is not exhaustive and some remarks are necessary:

- The choice between causal (physical) and statistical description of a natural process is mostly determined by mathematical ease and less by the deep physical background. E.g. the decay of radioactive elements is essentially random, but it is comfortable for practical use to describe it by a 1st-order deterministic differential equation. On the other hand, a flood is a simple deterministic process: the flow in a river is the sum of the affluents plus the water coming from rain or melting snow, if such is the case. It is, however, more convenient to describe the process by a multi-dimensional auto-regression model (see below).
- The causal (physical, deterministic) models involve constants, which must be determined by experiments, e.g. the constant 9.81 m/s^2 characterizing the Earth´s gravity potential. Each experiment involves errors which are random, if the experiment is correct. Thus every deterministic prediction contains a random component.

1.3. Prediction in chaotic systems

Chaos may occur when the behavior of a deterministic system is very sensible to initial conditions. It may be shown by numerical modelling, that the motion of three bodies is in a general case chaotic, i.e. unpredictable. This applies still more for the motion of many bodies, such as the planetary system. It is not possible to forecast the impact of meteorites; even the long-term stability of planetary orbits seems to be under question. However, sensitivity to initial conditions does not fully determine the chaotic systems: e.g. the function $f(x)=\exp(x)$ is very sensitive to initial conditions, but by no means chaotic. The chaotic behaviour is implied by topological mixing: in a chaotic system, any open set in the phase space will eventually overlap with any other open set.

Probably the most important partial differential equation in Earth Science is the Navier-Stokes equation

$$\partial v_i/\partial t + (\partial v_i/\partial x_j)v_j = G_i - (\partial p/\partial x_i)/\rho + (\mu/\rho)\,\Delta v_i\,,$$

with symbols as usual (v – velocity, p – pressure, x – space coordinates, G – gravity force etc). This equation applies for the motion of water in rivers and oceans as well as the motion of air masses. The non-linearity of the equation causes the impossibility of a long-time weather prediction.

An award of one million US dollars has been promised to whom finds the analytic solution to this equation. The rough features of the solution may be guessed when observing the motion of running water: for low velocities, the solution approaches the potential flow, while for high velocities, the motion is turbulent and may be (probably) described by a fractal function.

For practical use in hydromechanics and civil engineering, the Navier-Stokes partial differential equation used to be replaced by algebraic approximations, e.g. the Chézy equation

$$v = C\,\sqrt{(Ri)}\,,$$

where C is a material constant determined by the roughness of the channel or river bed, R is the „hydraulic radius“ (surface of the cross-section divided by the length of the bottom and the banks in the cross-section) and i is the inclination. It is interesting to compare the Navier-Stokes equation against the Chézy equation and thus to identify the loss of information resulting from the algebraic simplification:

- the Chézy equation cannot result in chaotic motion, and thus cannot predict the turbulence,
- the Chézy equation predicts only the mean velocity, not the distribution of velocity in the cross-section,
- the Chézy equation omits the vorticity and thus cannot serve as a basis for considerations concerning interaction between running water and surrounding soil or rock.

2. Statistical models of prediction

3.1. Stationary time series

Autoregression models: AR(1): $X_t = \sum_{k=0}^{\infty} c_k Y_{t-k}\,,$

Y is a white noise, $c_k = a^k$, abs (a) ‹ 1,

$X_t - a X_{t-1} = Y_t.$

Operator of delay B:
$(1\text{-}aB)\ X_t = Y_t$,
$X_t = (1\text{-}aB)^{-1}\ Y_t$.
Prediction in AR(1): $X_{t+m} = a^{m+1}\ X_{t-1}$ for m=0,1,.... (Markov chain).

AR(n): $X_t = \Sigma\ c_k\ Y_{t-k}$,
$\Sigma\ abs^2\ (c_k) < \infty$
$a_0\ X_t + a_1\ X_{t-1} + ... + a_n\ X_{t-n} = Y_t$.

Moving average models: MA (n): $X_t = \sum_{k=0}^{\infty} a_k\ Y_{t-k}$

MA (1): $a_0 = 1$, $a_1 = -b$
$X_t = Y_t - b\ Y_{t-1}$
$X_t = (1 - bB)\ Y_t$
Prediction in MA (1):

$$X_t = -\sum_{k=0}^{n} b^k\ X_{t-k} \text{ (one step ahead),}$$

$X_t = 0$ (two or more steps)

ARMA (m,n): combination of AR, MA.

3.2. Non – stationary time series and other generalizations

Random walk: $X_t = X_{t-1} + Y_t$ (generalization of AR(1) to the case a = 1)
$(1\text{-}B)\ X_t = Y_t$

Generalization FI (fractional integrated process):
$(1\text{-}B)^d\ X_t = Y_t$, stationary and invertible for $-0.5 < d < 0.5$;
it corresponds to long – memory processes (Hurst 1957, Granger – Joyeux 1980).

Other models currently in use: ARIMA (auto-regression integrated moving average), SAR (seasonal auto – regression), SMA (seasonal moving average), SARMA (seasonal moving average), SARIMA (seasonal integrated mixed model), ARFIMA (fractional integrated mixed process), SETAR (self – exciting threshold auto regressive), STAR (smooth transition autoregressive), MSW (Markov switching) etc. Most of them have been used in the analysis of economic time – series (Arlt – Arltová 2007).

In general, the use of sophisticated models is limited by the necessity of long series of reliable data.

3.3. Trend analysis

Least-squares method (linear regression)
Advantages:

- unique and explicit solution,
- easy determination of confidence intervals,
- easy generalization to a polynomial trend,
- possible generalization to non-linear regression by means of the singular value decomposition of the regression matrix

Disadvantages: enhances the influence of remote values.

LAD (least absolute deviations)
Advantages: less sensitivity to remote values,
Disadvantages:
- more complicated calculations,
- solution is not unique.

3.4. Mean frequency of occurrence

The method is often used in practice, in fact it is the only method of prediction which is used on a large scale. The method does not predict a single event, but the number of events (e.g. earthquakes of a given intensity) during a time interval. The method is very robust, but it requires long time series. To define the value of a 100-year flood, data of at least 1000 years series would have to be registered, which can hardly be satisfied. The lack of necessary information may be overcome in two ways:
- filling the missing data from literature and indirect indications,
- extrapolating the frequency of occurrence on the basis of a given distribution (e.g. the Weibull´s distribution or Fréchet distribution or Gumbel´s distribution are used, see Brázdil – Kirchner et al. 2007).

4.Several topics

4.1. Earthquake forecast (wikipedia, 2011)

- emissions of radon,
- seismic electric signals,
- foreshock prediction: successful prediction of the Haicheng earthquake in 1975. 50% of major earthquakes are preceded by foreshocks, but only 5-10% of small earthquakes are foreshocks, followed by major earthquakes.
- pattern theories (improvement of the preceding method),
- fractoluminiscence (only theoretical proposal),
- satellite magnetometry,
- animal behaviour,
- seismic hazard assessment: probability that an earthquake of a given size will affect a region during a given time interval.

In general, none of the methods of prediction of an individual earthquake is satisfactory, if both the error of the 1st and 2nd kind is considered.

4.2. Weather forecast and climate change

Simultaneous numerical solution of the Navier-Stokes equation, the equation of continuity and the thermodynamic energy equation (1^{st} law of thermodynamics). Problems: influence of clouds, non-linearity of the equations. The forecast seems to be reliable for several days, but long-term forecast (climate change) predicts the climate for many years. The IPCC model from 1990 signalized for the year 2000 the global average temperature 1^0C higher than the reality (Lomborg, 2001). Subsequently, the program was corrected, but too many scenarios are currently presented.

4.3. Floods

Long-time prediction: only frequency of occurrence.
Short-time prediction: multidimensional auto-regression AR(1) may be applied (Anděl 1976)

$$\mathbf{X}_{s+1} = \mathbf{U}\,\mathbf{X}_s,$$

where the vectors **X** are the flows in the main river and its tributaries and the coefficient matrix **U** may be determined from the past values by the least-squares method or by the Bayesian method. This prediction works well for small rivers, but for large streams it is in practice impossible to get sufficiently reliable input data.

4.4. Water balance

My working team has been busy with the water balance of the Morava river basin since the last 10 years (Chyba et al., 2003-2011). The basin of the Morava river belongs to the river system of the Danube river; it is situated in the SE part of the Czech Republic and with its surface of 20000 square km, it covers about ¼ of the Czech Republic. The water balance involves the „book-keeping“ of the quantity and quality of the surface – and groundwater, the quantity and quality of the withdrawn and discharged water in individual streams and hydrogeological regions, both for the present state and the future perspective.

The prediction is necessary for the „water policy“, e.g. planning if dams or sewage plants, permits to withdraw fresh water or to discharge wastewater etc. The actual and future demands are compared with the capacities of the sources and recipients.

In Fig. 1, the time-series of the withdrawal of surface water (dashed line) and groundwater (full line) in the years 2001-2010 is presented along with the prediction for the year 2011, which is based on the AR(1) model.

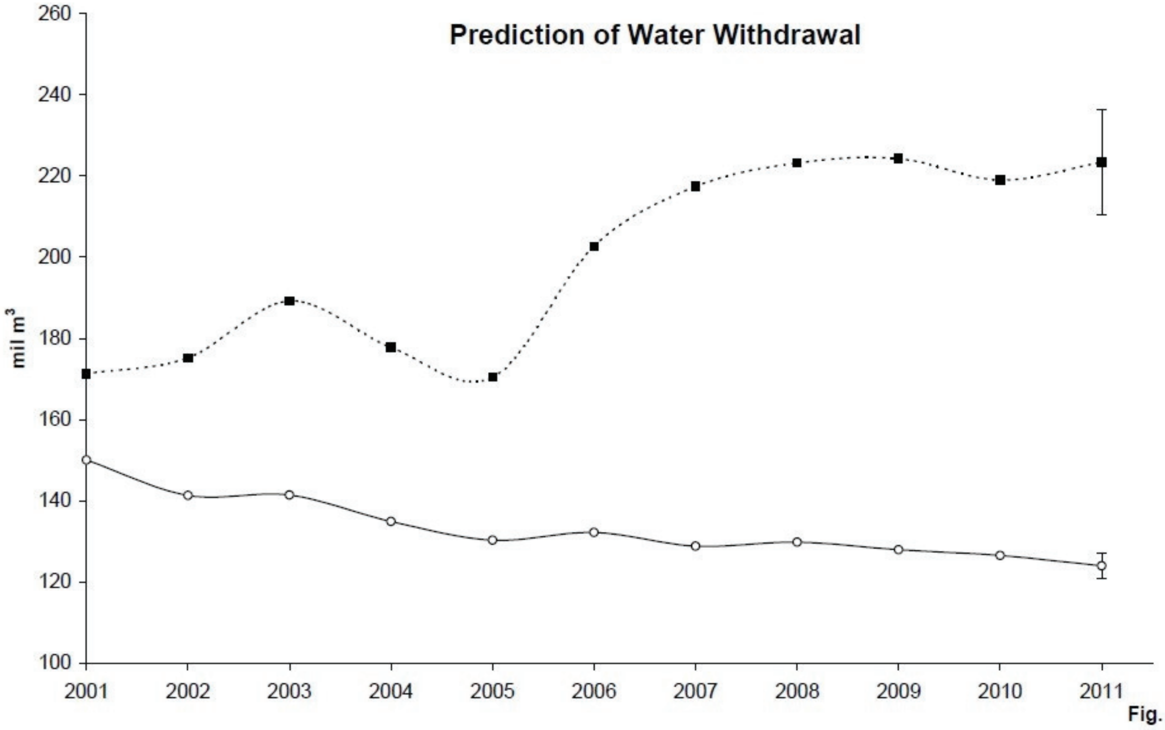

Fig. 1

In spite of the relatively flat curves, it is obvious that the water balance may be strongly disturbed by economic boom or crisis. This applies still more, when the balance is considered in detail for individual streams. The curve of withdrawal of the groundwater is much smoother than that of the surface water. This is so because groundwater is reserved as drinking water, and thus is less exposed to the changes of industrial needs.

3. Discussion

Any scientific prediction must be based on the knowledge of the history and the assumption that some features (statistical distribution, trend, physical laws controlling the process etc) remain constant. The most difficult problem is the definition of reliable and homogeneous data series. Any improvement of methodology implies non-homogeneity of data. E.g. the construction of water balance has been done repeatedly for many years in the Czech Republic, but minor changes have been made such as increasing the number of monitoring wells, definition of hydrogeologic regions, improvements in chemical analysis of specimens etc. Improvements in methodology must be done very carefully, so as not to disturb the homogeneity of data.

Another problem concerns the discrimination, i. e. the simultaneous reduction of both the errors of the 1st and 2nd kind. It was mentioned above that the foreshock method signalized the Haicheng earthquake in 1975. On the other hand, a major earthquake was predicted in Lima in 1980, which did not occur. Obviously, a false alarm causes important economic losses which could be claimed against the author of the prediction.

Analysis of frequency of occurrence is the only prediction method which seems to be appropriate for a practical use on a large scale. The method does not predict individual events, and thus is not vulnerable from a legal point of view. It is particularly convenient in allocating investments, e.g. protection of settlements against floods is allocated according to the criterion of effectivity, which is defined by the index

$P = \sum$probable damage/costs.

However, this criterion may be subject to criticism, because in enhances investments into the protection of large settlements. Inhabitants of small villages are discriminated and the procedure which seems to be correct from the economic point of view violates the principle of equal treatment.

4. Conclusion

I resumed in the present paper the most often used methods of prediction in Earth Sciences along with some examples of application. It appears that processes controlling natural phenomena are mostly non-linear, thus giving rise to chaos and making a long-term prediction impossible. Deep study of non-linear partial differential equations seems to be necessary so as we might better understand the limits of prediction.

However, the most difficult problems are not in mathematics, but in ethics and law. It is necessary to define the scope of responsibility for a wrong prediction. Furthermore it is necessary to equilibrate the economic point of view (maximum effect at a given cost) with the ethical and legal principle of equal treatment.

References

Anděl, J. (1976): *Statistická analýza časových řad (Statistical analysis of time series).* SNTL, Praha

Anděl, J. (2003): *Matematika náhody* (Mathematics of hazard). Matfyzpress, Praha, ISBN 80-86732-07-X.

Arlt, J. – Arltová, M. (2007): *Ekonomické časové řady* (Economic Time Series). Grada, Praha, ISBN 978-80-247-1319-9.

Brázdil, R. – Kirchner, K. et al. (2007): *Vybrané přírodní extrémy a jejeich dopady na Moravě a ve Slezsku* (Selected Natural Phenomena and their Impacts in Moravia and Silesia). Masaryk University, Brno, ISBN 978-80-210-4173-8.

Chyba, J. et al.: *Vodohospodářská bilance povodí Moravy* (Water Balance of the Morava River Basin) 2002-2010, Povodí Moravy, s.p., Brno.

Granger, C.W.J. – Joyeux, R. (1980): *An introduction to Long Memory Time Series Models and Fractional Differencing.* Journal of Time Series Analysis1, 15-29.

Hurst,H.E.(1957):*A suggested Statistical Model of some Time Series which occur in Nature.* Nature, v. 180, p. 494.

Lomborg, B.(2001): *The Sceptical Environmentalist.* Cambridge University Press.

http://en.wikipedia.org/wiki/Earthquake_prediction.