# The uncertainty enabled model web (UncertWeb)

Edzer Pebesma[1], Dan Cornford[2], Stefano Nativi[3], and Christoph Stasch[1]

[1] Institute for geoinformatics, University of Münster, Germany, Weseler Strasse 253, 48151 Münster, DE; edzer.pebesma@uni-muenster.de
[2] Computer Science and NCRG, Aston University, Birmingham B4 7ET, UK
[3] CNR and University of Firenze, Italy

**Abstract.** UncertWeb is a European research project running from 2010–2013 that will realize the uncertainty enabled model web. The assumption is that data services, in order to be useful, need to provide information about the accuracy or uncertainty of the data in a machine-readable form. Models taking these data as imput should understand this and propagate errors through model computations, and quantify and communicate errors or uncertainties generated by the model approximations. The project will develop technology to realize this and provide demonstration case studies.

**Keywords:** interoperability, uncertainty, OGC, workflows, geostatistics

## 1 Introduction

Environmental models are often complex. They need to represent space and time in a discrete and referenced way, and typically need to cope with combinations of points, lines, polygons, or regular grids, and potentially irregular points in time (regular or irregular) or time intervals. For example, an air quality model may accept sources (emissions) from areas (agricultural sources), lines (representing traffic along streets) and points (e.g. construction sites or factories). It may discretize space as a regular grid in the horizontal direction and use vertical layers of different thickness in order to solve the continuity equation that describes the change over time of the concentration of a component resulting from transport, chemistry, dry and wet deposition, and emissions. As an input it may already require wind velocity fields generated by a numerical weather prediction model, or a Monte Carlo sample (ensemble) of these that summarize our limited knowledge about the atmosphere.

According to [1],

> *The Model Web is a concept for a dynamic network of computer models that, together, can answer more questions than the individual models operating alone [...]. It is based on a philosophy that encourages modellers to provide access to their models and model outputs via standard "web services" [...], which makes it easier for the models to exchange information.*

Model inputs may be outputs from other models, but may also be sensor data, e.g. temperature sensor readings. To function well, the model web should provide standard web interfaces to allow both access to model resources and data resources. The seemingly sharp distinction between sensor readings and model outputs might be not that sharp: sensors need to do some processing to translate an electronic, analogue signal into a binary representation of a number reflecting temperature in some unit, and models cannot distinguish between measured or modelled values when they read them as input.

A major motivation to convert models and data sources into web services (or encapsulate models in web service interfaces) that are standardized, is that it allows an easier coupling of model and data resources, and allows the exchange of resources to evaluate robustness to particular choices for model representations or data resources. When access to model and data resources is open, a further motivation is that it democratizes access to scientific data and computation, and enhances the possibility to reproduce scientific research, and hence the transperency of scientific activity.

In addition to openness about the procedures used to reach a given result, a second pilar to obtaining transparency of scientific research is openness about the limits of knowledge obtained. Measurements have limited accuracy. Models are approximations to reality. Models using measurements result in values that are not identical to the "reality" modelled. The most common way to represent the degree of knowledge of a given quantity is by representing quantities through probability theory[4].

The UncertWeb project (2010-2013) will realize instances of the model web, and set it up in such a way that individual components of it (i) can understand inputs that are encoded as uncertain quantities, for example described using probability distributions rather than fixed values, (ii) can propagate errors through the model computations, and (iii) can output errors generated as a result of model *approximation*. At the time of writing this paper the project has barely started, but a number of technological challenges that will be faced, as well as the case studies that will be addressed, will be given.

## 2 Chaining environmental model web services

Web services need to advertise what they assume as inputs and outputs, in order for machines to be able to check that they meet the assumptions made. The industry standard for this is to us the SOAP/WSDL protocol. The Open Geospatial Consortium (OGC, [2]), a standardisation body dedicated to geospatial data handling, has made assumptions about OGC-compliant web services that do not match the SOAP protocol. Using OGC Web Services (OWS) has

[4] Probability theory is traditionally used to represent aleatory uncertainty (variability) and epistemic uncertainty (lack of knowledge). An alternative that models *concepts* as vague (sometimes called ontological uncertainty) rather than crisp in a quantitative fashion is to use *fuzzy set theory*, although some argue that this can be addressed within a subjective Bayesian perspective.

the advantage that software in the geospatial domain understands them (read: clients exist that can work with their output), but has the disadvantage that standard interactive, visual tools for web service chaining such as Kepler, Taverna and BPEL enginges [4,3,6] for workflow orchestration do not work without a lot of additional effort. Within UncertWeb we are exploring a number of ways of integrating OWS with the SOAP/WSDL world.

### 2.1 O&M, WPS, profiling

Model webs, consisting of environmental models that are used for nowcasting typically integrate current sensor data in the workflow. Sensor data can be provided over the OGC Sensor Observation Service, which yields either a SensorML document or a document conforming to the *Observations and Measurements* (short: O&M) conceptual model. Because observations and measurements can, in principle, be anything from a temperature reading to a video stream to a text message telling "It's hot in here", and time and space descriptions can have many forms, this standard has to cope with a potentially huge number of possible data representations. This makes it hard to predict whether an arbitrary O&M compliant document will be accepted by a particular service. To limit the scope of possibilities, so called *profiles* for particular applications or application domains, can be developed as a workaround. This is important to allow developers to support only the aspects of O&M that are required in the given application context. Attemping to support the complete conceptual model would be beyond the scope of any existing systems.

An OGC standard for a service that could execute a model is the Web Processing Service (WPS). This service puts no constraints to the type of process that can be executed, nor what the input should and output will look like. Again, the advice is to develop a *profile* to describe the constraints on this in a machine-readable way.

### 2.2 UncertML

Error propagagion, and in particular error propagation by Monte Carlo simulation is not complicated in nature—instead of running a model once it is simply run, say, 1000 times with inputs sampled from the distribution that reflects the uncertainty of this particular input. The tricky bit is to organize the set of inputs and outputs when they have complicated structure (e.g. are a time series of maps), and to derive relevant information from them.

One layer of information needed is that of semantics of the uncertainty: which names do we give particular properties, such as the 10th realisation in a Monte Carlo sample, the mean and variance, or the 95-percentile. UncertML [7] is the starting point for a markup language to encode uncertain information. UncertWeb will improve and test this first version.In particular within UncertWeb, UncertML will be separated into a simple conceptual model, a controlled vocabulary and series of encodings.

### 2.3 Discovery

UncertWeb discovery functionalities will consider environmental models descriptions expressed as specific metadata model profiles. These will include the extra information needed to discover and chain models according to significant use cases. The proposed solutions will consider the existing and ongoing specifications from relevant initiatives (e.g. GEOSS, GMES, INSPIRE, etc.). The profiles will be based on well-accepted international standards, where appropriate.

As far as discovery and access are concerned, presently UncertWeb is investigating the following challenges:

a Uncertainty support for Scientific Data Types, with particular reference to the next data model and encoding languages: Geography Mark-up Language (GML), Obeservation&Measurement (O&M), and Common-Data-Model (CDM) / netCDF / ncML;
b An extended SOA approach addressing interoperability for uncertainty-enabled sharing of environmental resources. A SOA brokering approach was successfully experimented in several GEO/GEOSS initiatives and FP7 projects (e.g. GEO AIP, EuroGEOSS project, etc.) to reduce/solve interoperability issues on discovery, access, and semantics functionalities.
c Support of uncertainty in discovery services; this concerns: (i) possible uncertainty queryables, e.g. relative/average error on a data coverage; (ii) uncertainty accommodation in standard metadata models, e.g. ISO 19115, INSPIRE Metadata profile, etc.

### 2.4 Service chaining

As for chaining and publication methodology and services, presently UncertWeb is defining the standard baseline for chaining, including standard services (e.g. OGC services and ISO models) as well as communities of practice specification and special interoperability arrangements, e.g. from the biodiversity community.

Besides, in keeping with the Model Web principles, UncertWeb is investigating a suitable approach for chaining uncertainty-enabled services, starting with data and processing services. Different possible solutions (i.e. legacy frameworks such as Kepler/Taverna/JOpera [4,3,5]; standard environments such as BPEL [6]; Composition-as-a-Service, and Mash-ups that could include Google maps) are compared evaluating their infrastructure complexity, entry level barrier, flexibility, and portability.

An extended SoA approach, applied to the data service domain, promises to facilitate workflows chaining and accomplish brokered data access for not directly interoperable nodes. In fact, this approach could also enable the transformation/mediation of uncertainty related metadata description and uncertain scientific data-types.

### 2.5 Methods and tools for uncertainty management

Generic technology for building uncertainty enabled model webs is only partially available, and UncertWeb will need to develop some more. Among these

are improved Monte Carlo methods, needed to reduce the number of model runs required when model runs are computationally expensive. Other developments involve expert elicitation, a method where experts opinions are used, and combined to obtain quantitative uncertainty measures for quantities for which this is hard or impossible to obtain experimentally. Spatial, temporal or spatio-temporal aggregation or disaggregation is often needed to obtain information at a finer or coarser spatial resolution, and has consequences for uncertainty: the average value for a country is easier to estimate than the value for each point. The value, limitations and requirements to Monte Carlo simulation methods for error propagation will also be studied in the context of complex models with high-dimensional parameter spaces. Finally, visualisation of uncertain information [10] is an active area of research, and is of particular importance to communicate information to end users, who could be domain experts or members of the public. Usability testing is foreseen to evaluate different alternatives.

## 3 Case studies

### 3.1 Biodiversity modelling

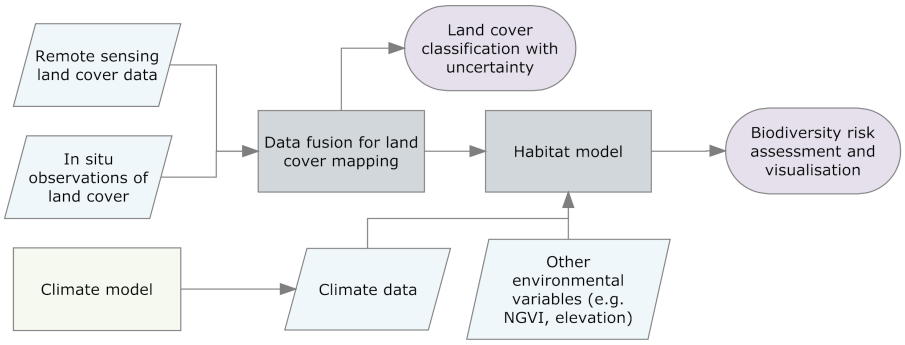


**Fig. 1.** Work flow for biodiversity modelling

The system developed in this case study (Fig 1) will enable land cover data to be efficiently exchanged and validated, prior to being used to measure (mainly anthropogenic) changes to protected zones. The model server that will be put in place to compute and serve interoperable habitat information will benefit from validated information, and the uncertainty framework developed in UncertWeb will further allow us to quantify uncertainties in habitat changes. The coupling in a Service Oriented Architecture (SOA) environment of this habitat model server with climate change model servers will provide powerful new mechanisms to policy makers and the scientific community for performing multi-scale analyses of biodiversity-related problems and for forecasting ecological change.

### 3.2 Food chain modelling

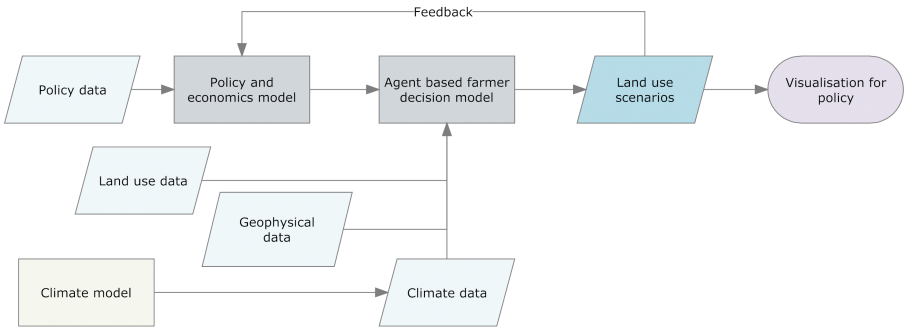


**Fig. 2.** Work flow for food chain modelling

The prototype systems developed in biodiversity modelling work package (Fig. 2) within UncertWeb will enable policy makers to integrate computer models that simulate land-use and its change under different economic and climatic scenarios. An important advance is the ability to propagate uncertainty through the model Web: instead of focusing on highly constrained scenarios (as in the FP6 LUMOCAP project, http://www.riks.nl/projects/LUMOCAP), we can consider the impacts of uncertainty at the policy level. The model Web will also allow us to consider other facets of the land-use problem, and we could open up the possibility of using some more detailed environmental models (for examples, see [11]).

### 3.3 Air quality modelling

The application of uncertainty-enabled forecasts for air quality (Fig. 3) is a logical and inevitable extension of numerical weather prediction ensemble forecasts. Whilst current developments mainly focus on the regional scale, the concept is just as applicable, and indeed more useful, for local authorities on the urban scale. It is expected that in the near future such ensemble forecasts will be commonly available, and so it is essential to develop a communications system suitable for dealing with such information.

### 3.4 Human activity modelling

Activity-based modelling of individual movements (Fig. 4) is an area that is just entering the operational phase, because of its complexity, demands on input data, and runtime requirements. The quantification of the prediction errors which result from model approximation and input uncertainties is a novel application

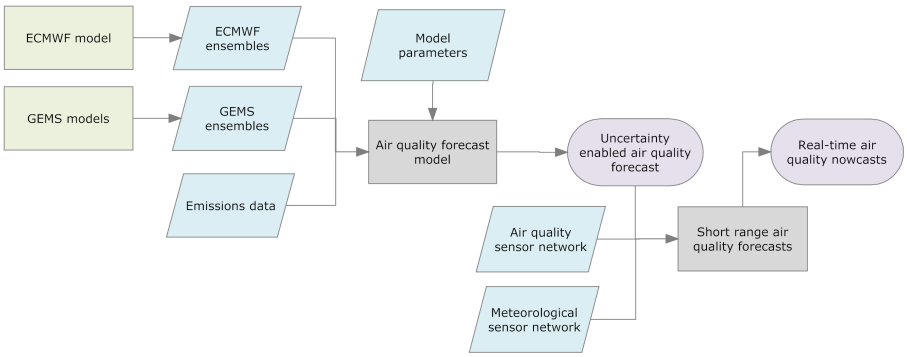


**Fig. 3.** Work flow for air quality modelling

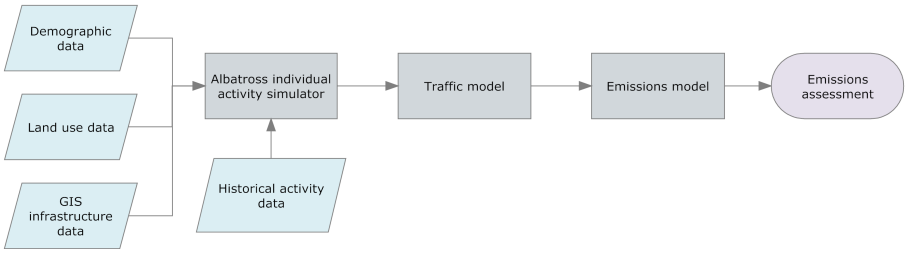


**Fig. 4.** Work flow for human activity modelling

area, but is required in order to acknowledge and manage the limited reliability of outputs from realistic models.

### 3.5 Integrating air quality and human activity modelling

Successful integration of the two case studies on air quality forecasting and individual activity modelling (Fig. 5) will be an important proof of concept for the uncertainty-enabled model Web. The application of human activity modelling is the first of its kind in Europe, but its predictions are subject to uncertainty. Using these uncertain predictions to feed an air quality model is a challenging and innovative approach, and will allow assessment of the input (predicted emissions from traffic) as well as assessment of exposure based on realistic allocation of individuals in space and time.

## 4 Discussion

UncertWeb is an ambitious program that has the potential to demonstrate that environmental model webs work, and can sensibly cope with uncertainty in a

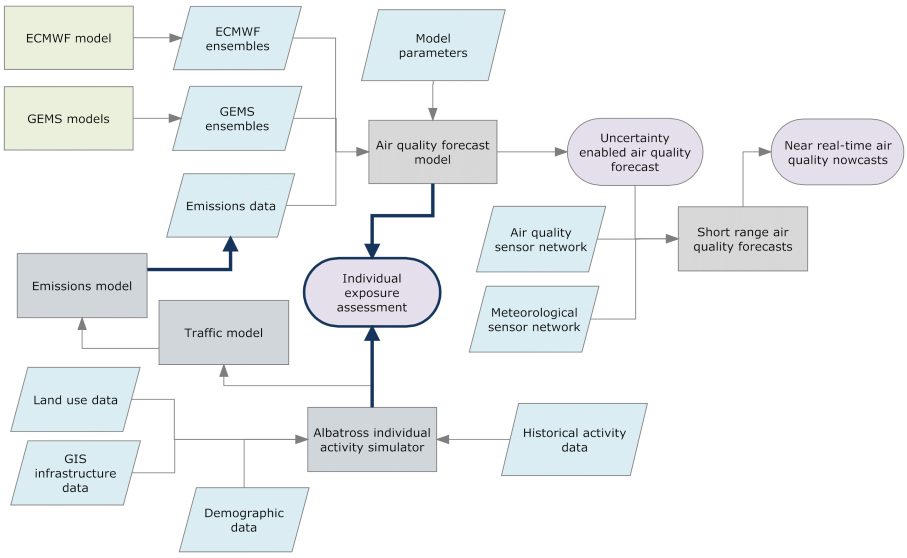


**Fig. 5.** Work flow for integrating air quality and human activity modelling

quantitative way. A number of technological issues to realize this will have to be tackled, solved or circumvented, in four very diverse case studies a number of cutting edge challenges, including that of linking physical atmospheric modelling (air quality) with social, agent-based human activity modelling while taking into account the major sources of uncertainty. In particular we aim to address the usability of the resulting system, providing tools to enable people to assess, manage and visualise uncertainties in chains of web services.

## Acknowledgements

This work has been funded by the European Commission, under the Seventh Framework Programme, by Contract No. 248488 with the DG INFSO. The views expressed herein are those of the authors and are not necessarily those of the European Commission. More information on UncertWeb and UncertML can be found on http://www.uncertweb.org/ and http://www.uncertml.org/, respectively.

## References

1. Geller, G.N. and Melton, F. Looking Forward: Applying an Ecological Model Web to assess impacts of climate change. Biodiversity 9 (3&4), 79–83.
2. http://www.opengeospatial.org/

3. https://kepler-project.org/
4. http://www.taverna.org.uk/
5. http://www.jopera.org/
6. Tan W., Missier P., Madduri R., Foster I. (2009). Building scientific workflow with Taverna and BPEL: A comparative study in caGRID. In: Proceedings. 4th International workshop on Engineering Service-Oriented applications (WESOA), pp. 118129.
7. Williams, M., Cornford, D., Bastin, L. and Pebesma., E. (2009). Uncertainty Markup Language (UncertML). Discussion Paper 08-122r1. Open Geospatial Consortium. Available from http://portal.opengeospatial.org/files/?artifact_id=33234.
8. OGC (2007a). Sensor Observation Service. Available from http://portal.opengeospatial.org/files/?artifact_id=26667.
9. OGC (2007b). OpenGIS Web Processing Service. Available from http://portal.opengeospatial.org/files/?artifact_id=24151.
10. E.J. Pebesma, K. de Jong, D.J. Briggs (2007). Visualising uncertain spatial and spatio-temporal data under different scenarios: an air quality example. International Journal of GIS, 21, 515527
11. EEA (2008). Modelling environmental change in Europe: towards a model inventory (SEIS/Forward). EEA Technical report No 11/2008.