Uncertainty Propagation in the Model Web. A Case Study with eHabitat

J. Skøien*[a], P. Truong[b], G. Dubois[a], D. Cornford[c], G.B.M. Heuvelink[b], G. Geller[d]

[a] European Commission, Joint Research Centre, Institute for Environment and Sustainability, Ispra, 21027(VA), Italy – (jon.skoien, gregoire.dubois) @jrc.ec.europa,eu
[b] Wageningen University, Land Dynamics Group, Wageningen, the Netherlands – (phuong.truong, gerard.heuvelink) @wur.nl
[c] NCRG and Computer Science, Aston University, Birmingham, UK - d.cornford@aston.ac.uk
[d] NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA – gary.n.geller@jpl.nasa.gov

eHabitat is a Web Processing Service (WPS) designed to compute the likelihood of finding ecosystems with equal properties. Inputs to the WPS, typically thematic geospatial "layers", can be discovered using standardised catalogues, and the outputs tailored to specific end user needs. Because these layers can range from geophysical data captured through remote sensing to socio-economical indicators, eHabitat is exposed to a broad range of different types and levels of uncertainties. Potentially chained to other services to perform ecological forecasting, for example, eHabitat would be an additional component further propagating uncertainties from a potentially long chain of model services. This integration of complex resources increases the challenges in dealing with uncertainty. For such a system, as envisaged by initiatives such as the "Model Web" from the Group on Earth Observations, to be used for policy or decision making, users must be provided with information on the quality of the outputs since all system components will be subject to uncertainty. UncertWeb will create the Uncertainty-Enabled Model Web by promoting interoperability between data and models with quantified uncertainty, building on existing open, international standards. It is the objective of this paper to illustrate a few key ideas behind UncertWeb using eHabitat to discuss the main types of uncertainties the WPS has to deal with and to present the benefits of the use of the UncertWeb framework

Keywords: Uncertainties, SOA, ecological forecasting, Web Processing Services, Model Web

1. INTRODUCTION

Despite being an old science, ecology lags behind newer sciences such as the field of Earth System Sciences when it comes to integrating different disciplines and methodologies. The complexity is large and the amount of data and variety of formats is considerable. However, with the threats of climate change, there are now a range of initiatives to gradually integrate models for a better understanding of the complexity of ecosystems. The Group on Earth Observations Biodiversity Observation Network (GEO BON), for example, is starting to address these issues by facilitating the harmonization of existing biodiversity observation systems. By linking services together in a "Model Web" in an interoperable manner (Geller and Turner, 2007), it will be possible to reuse results and methodologies across disciplines. In the framework of the development of the Digital Observatory for Protected Areas (DOPA), a biodiversity information system currently developed as a set of interoperable web services at the Joint Research Centre of the European Commission in collaboration with other international organizations, currently including the Global Biodiversity Information Facility (GBIF), UNEP-World Conservation Monitoring Centre (WCMC), Birdlife International and the Royal Society for the Protection of Birds (RSPB), a Web Processing Service (WPS) for modelling probabilities to find similar ecosystems has been designed. The WPS-based system, called eHabitat (Dubois et al, 2011), is designed to become a basic component of the ecological Model Web that allows its functionalities to be chained with other modelling web services (e.g. climate change). Inputs to the WPS, typically thematic geospatial "layers", can be discovered using standardised catalogues, and the outputs tailored to specific end user needs. A schematic showing the information flow within eHabitat and a few possible links with other modelling services, like those proposed in the various initiatives of GEOSS, the Global Earth Observation System of Systems, is summarized in Figure 1.

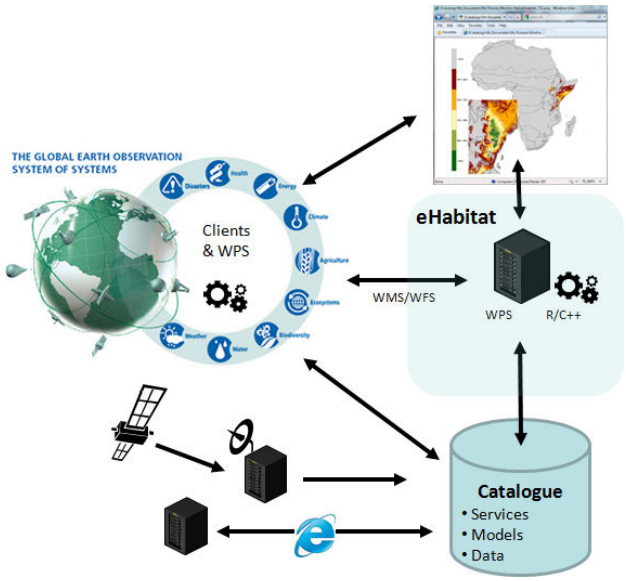

Figure 1 Data flow within the eHabitat web processing service. eHabitat computes probabilities of finding ecosystems that are similar to those found in a reference area (e.g. a protected area), by combining data from different sources.

Because these layers can range from geophysical data captured through remote sensing to socio-economic indicators, eHabitat is exposed to a broad range of different types and levels of uncertainties, which are inevitably propagated through the service (see e.g. Heuvelink, 1998). The UncertWeb project (www.uncertweb.org) will create the Uncertainty-Enabled Model Web by promoting interoperability between data and models with quantified uncertainty, and building on existing open, international standards. UncertWeb will thus develop open source implementations of encoding standards, service interface profiles, discovery and chaining mechanisms, and generic tools to realize a "Model Web" taking uncertainty in data and models into account.

2. UNCERTAINTIES AND THE MODEL WEB

2.1 The Model Web

"*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*" (Geller and Melton, 2008). In other words, the Model Web is composed of a large set of models that are exposed as Web Services and chained in an interoperable way. The users of one model exposed as a web service may not notice that the model makes requests to other web services to be able to handle the original users requests. There are several purposes for this chaining. Each modeller can specialise and refine their own models, for which they are the experts. Their knowledge of the requested models can be limited to the specifications of the models input, output and usability. This will eliminate the need for downloading and installation of a range of models that for which the modeller has less expertise. It will also eliminate the need for updating these models because the web services will be updated as soon as the model owner identifies and fixes bugs and security holes. Linking the models to the web service interfaces of data providers will further make it possible to predict complex relationships in real time with minimum effort. Lastly, as the Model Web becomes more mature, it will be easier to exchange components of the chain with competitive models, either for comparisons between models, or for ensemble predictions. One challenge in making the models interoperable is that the interface of the models should be as generic as possible, using open, commonly used and accepted standards. Closed, particularly complex or uncommon formats will restrict the use of a particular model in a model chain.

2.2 The Uncertainty-Enabled Model Web

In the attempt to simplify a model, uncertainty is often the first victim. Uncertainty handling will for many models dramatically increase computation time, and the size of inputs and outputs will in general double or worse. This is not only a challenge within the Model Web. Many desktop models ignore the uncertainty of inputs and outputs, and treat the results as certain. The risk here is that users will be overly confident on the produced results, and not able to distinguish between precise predictions and predictions that are close to qualified guesses. A result without its uncertainty is often of limited value. The need for uncertainty propagation and proper expression of the uncertainty of output increases with increased complexity of the model chain and for interoperable models in general. When it is prohibitive or impossible for the user to examine the intermediate results, it is even more important that the model chain is able to propagate uncertainty and return a result with quantified uncertainty.

UncertWeb will build on the Model Web concept and contribute to it by supporting accountable uncertainty representation and propagation. A range of different tools and extended standards are necessary to uncertainty-enable web services. First of all, UncertWeb will further develop UncertML (www.uncertml.org), which is an XML (Extensible Markup Language) encoding designed for encapsulating probabilistic uncertainties. This encoding is necessary for interoperable communication of uncertainty between web services. The flexibility of XML encodings is high, but for larger data sets such as spatial grids, UncertWeb will also contribute to extended standards for NetCDF (network Common Data Form).

There are two properties of a model exposed as a web service that are of particular importance when we want to uncertainty-enable the service itself:

- Whether the user or the model sending a request can control the input, or if the service provider has restricted the input to one or more particular data sets

- Whether the model itself is able to propagate uncertainty

Let uWPS be in this context an uncertainty-enabled Web Processing Service that is able to propagate the uncertainty of the input through the model, analytically or through a Monte Carlo approach using different realizations of the input. The uncertain result is then returned from the service as UncertML and/or netCDF. Let UWS be an Uncertainty Wrapper Service, which can uncertainty-enable a service that is not able to take uncertainty into account. This is possible for services where the request includes the input to the model. The UWS will convert the uncertain input into realizations and apply a Monte Carlo approach when calling the service. The UWS can be seen as an extra layer on top of available Web Services. For different combinations of the model properties, one can summarize as in Table A the ways to enable web services to handle uncertainties.

Table A. Different ways of uncertainty enabling a web service depending on the models ability to propagate uncertainty and the user's control of the input.

		Model propagates uncertainty	
		YES	NO
User controlled input	YES	uWPS/UWS	UWS
	NO	uWPS	Not possible

It is obviously not possible to propagate uncertainty from a Web Service with restricted input to which the user has no access, and which is not able to propagate uncertainties itself.

The service denoted as uWPS above will be a service that is fully able to propagate uncertainty inside the model. In this case the service interface would use the UncertML Application Programming Interface (API) that manages the communication of uncertainty between the service (UncertML encodings) and the model itself. The API will in this case provide translation functionality to convert between uncertainty representations where this is possible but also includes the potential for a stronger link between the uncertain input and the model, such as analytic uncertainty propagation or use of more complex descriptions of the uncertainty. The API can also convert from probabilistic uncertainty of the input to Monte Carlo simulations of the model. The advantage of including this in the API is that network traffic will be reduced (because it will no longer be necessary to pass realizations over the internet).

2.3 The uncertainties of probabilities of habitat similarity (PoHS)

eHabitat is a simple WPS which is used to predict the Probability of Habitat Similarity (PoHS) between a set of points or a polygon of interest and the surroundings. The term habitat should be taken in the broadest sense as it is usually species specific. In the current, prototype version, of the modelling tool, the user can supply the service with a set of environmental indicators (climate, DEM, vegetation variables) as raster maps and the boundaries of a protected area (PA). The service uses the Mahalanobis distance to estimate the probability of the surrounding areas being environmentally similar to the PA. This can both be done for current data and for modelled data (such as climate scenarios) for ecological forecasting.

The core process in eHabitat is the computation of the Mahalanobis distances D which are used as a measure of the similarity, see e.g. Farber and Kadmon (2002). For each pixel, one can compute D which is defined as the square root of

$$D^2=(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})$$

where $\mathbf{x}$ is the vector of data, $\mathbf{m}$ the vector of the mean values of the independent variables and $\mathbf{C}^{-1}$ the inverse covariance matrix of the independent variables.

The use of the inverse of the covariance matrix $\mathbf{C}^{-1}$ makes the Mahalanobis distance dimensionless, i.e. it is not affected by the different scales of the measurements. The use of the covariance matrix also reduces the joint effect of highly correlated variables on D. When the predictor variables used to generate the mean vector and covariance matrix are normally distributed, then D is distributed according to a χ^2-distribution with n-1 degrees of freedom, and so we can convert D into p-values. The p-values (or probability values) range from 0.0 representing no similarity to 1.0 for areas which are identical to the mean of the PA. If the predictor variables are not normally distributed, the conversion is still useful as it rescales the unbounded D values to a 0.0 to 1.0 range. The p-value can be seen as the probability that a pixel outside the investigated area has a similar set of indicators as the ones found for the selected PA, or of the probability that, under a given scenario, a pixel will in the future have a similar set of indicators. After calculating the Mahalanobis distance for all pixels, a probability map showing the PoHS is returned to the user.

This metric-based approach to describe and compare ecosystems is exposed to several types of uncertainty, stemming from uncertainties in the data provided and in the processing. First of all, there are uncertainties in the thematic layers used for calculation of the PoHS. These uncertainties will in general be spatially correlated, in the sense that errors in one pixel tend to be similar to errors in adjacent pixels. We can therefore in the simple case describe the errors of these raster layers by two variables. First, for every pixel we need to know the standard deviation of the uncertainty. This standard deviation is assumed to be spatially constant. Second, we describe the spatial correlation through a variogram of the standardized uncertainty, where the sill is set equal to one, and we are most interested in the range. If the variables are also cross-correlated, we additionally need the cross-variograms.

From these assumptions, the easiest way to propagate uncertainty is through a Monte Carlo approach. For each variable we create a realization of the spatially correlated unstandardized uncertainty, using an unconditional simulation approach available through e.g. `gstat` (Pebesma, 2004). A set of unconditional simulations is created from a zero-mean process and a variogram with sill equal to one and range equal to the correlation length of the variables. These simulations are multiplied by the standard deviation of the uncertainty for each pixel, and added to the variable itself. When variables are cross-correlated, the procedure needs to be slightly modified by defining cross-variograms and using a co-simulation approach. A relatively large set of simulations is created (100 in the case below), and the PoHS is computed for each simulation. The user or requesting service can then decide if the results of all simulations neeed be transmitted, or only summary statistics. If summary statistics are chosen, the PoHS estimated from the original data set is returned. As indicated above, for proper uncertainty propagation cross-variograms are necessary for simulations of realizations of the input data in eHabitat. Ignoring cross-correlations will give incorrect PoHS, as realisations of the correlated variables would not vary simultaneously as they should. Estimating cross-variograms of errors might be a challenge. Perhaps we might assume that observed errors are spatially correlated if the variables themselves present clear spatial correlation, which is often the case with remote sensing data.

Another type of uncertainty to be considered, which will not be taken into account in this paper, is the positional uncertainty of the polygon of reference, in this case the boundaries of the PA.

2.4 Uncertainty propagation with the eHabitat WPS

eHabitat is meant to be one of the building blocks of the Model Web. As the methodology can be used for a range of purposes, the input can be of different character. Although an extremely simple model in itself, the coupling of eHabitat with other services like a climate change model service allows for potentially unprecedented ecological monitoring and forecasting tools as illustrated in Figure 1. Published as a web service and coupled with simple discovery tools for input layers, the user can efficiently evaluate the effect of different combinations of layers (variables). In the context of UncertWeb, a Web Service can either be uncertainty-enabled by accepting and returning data sets with uncertainty, or it can be uncertainty-enabled through a UWS employing Monte Carlo sampling for the uncertain input data. eHabitat is designed as a service that can propagate the uncertainty itself due to the possible large data sets, but it is also used as a test case for the Uncertainty Wrapper Service.

3. TEST APPLICATION

As a test application, we have analysed at different time steps the probability of finding similar habitats for the Serengeti National Park in Tanzania under a climate change scenario. The park is 15 000 km^2 in size, which is half the size of Belgium. The analyses in this paper are significantly simplified to illustrate the presented methods and no policy conclusions should be derived from the results presented.

3.1 Data set

The park boundaries of the Serengeti National Park have been downloaded from the World Data Base on Protected areas (WCMC-IUCN, http://www.wdpa.org/). For the application presented in this paper, we characterize the habitat considering only two of the climatic variables of Holdridge's lifezones, i.e., biotemperature and annual precipitation. The biotemperature is the annually averaged temperature after replacing all temperatures below the freezing point with zero values. These variables were computed from climatic data from the WorldClim data base (Hijmans et al., 2005), available at http://www.worldclim.org. Monthly values of the current climate (from the years 1950-2000) have been interpolated to a global raster from several thousand climate stations around the world and include temperature and precipitation. The data set is available in different raster formats, from 30 arc-seconds (approximately 1×1 km close to the equator) to 10 arc-minutes. The data for the future climate comes from two different scenarios from three large scale general circulation models (Hadley Centre, CSIRO, CCCMA). The outputs from these models were downscaled to the same resolution as the interpolated data sets of the current climate, assuming that the spatial patterns will be similar in the future (Ramirez and Jarvis, 2010). Due to the size of the Serengeti National Park, we restricted the analyses to the 10 arc-minutes data set, and we only present results from the CCCMA model.

3.2 Data uncertainty

Unfortunately we did not have access to the uncertainties on the climatic data sets at the time of writing this paper. Whereas the true uncertainty is dependent on a range of factors, such as the local density of climatic stations, variability of elevation and vegetation, we have simplified these here to be able to show a proof of concept. For the two variables, we assume that the standard deviation of the variables is 0.03 times the value. In this way, the uncertainty is proportional to the value of the variable, which seems to be a reasonable assumption. We also assume that the uncertainty has a correlation length that is equal to the correlation length of the variable itself.

3.3 Results

For each pair of simulations (biotemperature and annual precipitation), we calculated the PoHS. Figure 2 shows the estimated PoHS from the original data (pHab) and from some simulations for the year 2050. There are substantial differences between the simulations and the original data – some regions are predicted to have relatively high similarities for all simulations, whereas other areas have high similarities for some simulations and lower for others. Note also that the park itself is predicted to have high similarity with its current conditions when using the original data, but that two of the simulations indicate that the similarity can be rather low.

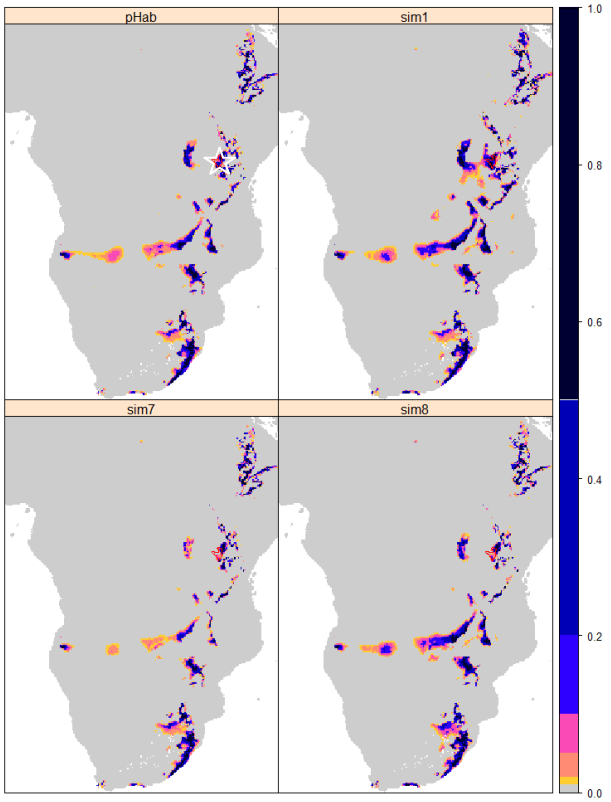

Figure 2. Estimated PoHS for 2020 using the original data and some realizations of the data. The location of the Serengeti Park in Tanzania is indicated with a white star in the upper left map.

4. CONCLUSIONS

We presented some of the first results and ideas for uncertainty propagation within eHabitat using the UncertWeb framework. The purpose of this paper was partly to present the concept, and partly to show the value of uncertainty propagation within the "Model Web". The possibility to use catalogues, discovery services and the results from other web services as input to eHabitat makes the service more flexible and useful both as a standalone service and chained with other services. The application example shows the value of uncertainty propagation in the calculation of PoHS. Some areas appear to have high PoHS values for the original data set and the different simulations, whereas other areas are simulated to have low PoHS for some realizations of the input data. An analysis based only on the original data could lead to wrong decisions regarding the future of the park and possible areas for replacement. Thus, quantified uncertainty has added value and should be included to minimize the risk of making poor decisions.

ACKNOWLEDGEMENTS

This work is partly funded by the European Commission, under the 7th Framework Programme, by the EuroGEOSS project funded by DG RTD and by the UncertWeb project funded by DG INFSO. The views expressed herein are those of the authors and are not necessarily those of the European Commission. More information on eHabitat & DOPA can be found at http://dopa.jrc.ec.europa.eu/

REFERENCES

G. Dubois, J. Skøien, J. de Jesus, S. Peedell, A. Hartley, S. Nativi , M. Santoro, G. Geller (2011). eHabitat: a contribution to the model web for habitat assessments and ecological forecasting. (*This Volume*)

O. Farber and R. Kadmon, (2002). Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance, *Ecological Modelling*, **160**:115-130.

G. Geller and F. Melton (2008), Looking forward: Applying an ecological model web to assess impacts of climate change, *Biodiversity*, **9**(3-4), 79-83.

G. Geller and W. Turner (2007), The Model Web: A concept for ecological forecasting, in *IEEE International Geoscience and Remote Sensing Symposium*, edited, Barcelona, Spain, 23-27 July.

A. Hartley, A. Nelson, P. Mayaux, and J.-M. Grégoire (2007), *The assessment of African protected areas*, JRC Scientific and Technical Reports, EUR 22780 EN, 70 pp, Office for Official Publications of the European Communities, Luxembourg.

G.B.M. Heuvelink (1998). *Error Propagation in Environmental Modelling with GIS* (Research Monographs in GIS). CRC Press, 144 p.

R. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis (2005), Very high resolution interpolated climate surfaces for global land areas, *International Journal of Climatology*, **25**, 1965-1978.

E.J. Pebesma (2004), Multivariable geostatistics in S: the gstat package, *Computers & Geosciences*, **30**, 683-691.

J. Ramirez and A. Jarvis (2010), *Disaggregation of Global Circulation Model Outputs*. Report from International Center for Tropical Agriculture, CIAT, Cali, Colombia. http://gisweb.ciat.cgiar.org/GCMPage/docs/Disaggregation-WP-02.pdf (accessed 11/1-2011)