

UncertWeb

The *Uncertainty Enabled Model Web*

SEVENTH FRAMEWORK PROGRAMME

THEME FP7-ICT-2009-4

ICT for Environmental Services and Climate Change Adaptation

Deliverable 4.1, 5.1, 6.1

Consolidated requirements for service chains within UncertWeb

Title of Deliverable	Consolidated requirements for service chains within UncertWeb
Deliverable reference number	D4.1, D5.1, D6.1
Related WP and Tasks	WP4 (Task 4.1), 5 (Task 5.1), 6 (Task 6.1)
Type of Document	Public Report
Authors	Dan Cornford, Gregoire Dubois, Jon Skoien, Jill Johnson, John-Paul Gosling, Bruce Denby
Date	29/07/2010
Version	4.0

Project coordinator

Dr. Dan Cornford
Aston University, United Kingdom
E-mail: d.cornford@aston.ac.uk

<http://www.uncertweb.org>

Revision History

Version	Date	Changes	Authors
0.0	23/02/2010	Initial template for all WPs	Dan Cornford
0.1	28/02/2010	Suggested revisions	Gerard Heuvelink
1.0	12/03/2010	Revisions following comments – sent to responsible partners	Dan Cornford
2.x	05/07/2010	Information added by responsible partners, often with several internal reviews.	Jon Skoien, Gregoire Dubois, Jill Johnson, John-Paul Gosling, Bruce Denby
3.0	07/07/2010	Consolidated the contributions of JRC, FERA and NILU	Dan Cornford
	21/07/2010	Review by UOM	Edzer Pebesma, Christoph Stasch
3.1	23/07/2010	Modifications by partners JRC, FERA	Jon Skoien, John-Paul Gosling
4.0	27/07/2010	Integration and final editing	Dan Cornford

Related task(s):

Task 4.1 User requirements for developing a tool for the validation of land cover data in African Protected Areas

This task will focus on defining the user requirements necessary to define the means to quickly visually validate high resolution multi-temporal satellite data (Landsat images) using ground truth data within and around African Protected areas. The user requirements will also address the reporting of uncertainties as well as the data discovery and exchange of data between the Web-based land-cover validation tool and the JRC database that is hosting the images. The requirements will be fed into WP1 and WP2 to inform the design of the Model Web encoding and interfaces.

Active partners: JRC, AST

Task 5.1 Defining the application domain requirements for service chaining

This task documents the models to be considered in the work package. This includes a description of each model's inputs and outputs, the computational complexity of the model and how the models link together. The interaction of socio-economic and policy models with the dynamic land-use model is of greatest importance. The final output will be a report that helps inform methodological developments in WP1-3.

Active partners: FERA, AST

Task 6.1 Establishment of data flows, standards and service chaining

This task will interact directly with WP1-3 and WP8 in order to effectively implement the protocols and standards for use in the uncertainty-enabled model Web. Current GMES services will be the main source of data for driving the ensemble forecasts. These source data will include a) ensemble weather forecasts available from ECMWF through GMES; b) ensemble regional scale air quality forecasts available from 7 models through the GEMS project; c) local near-real time monitoring data, currently available at NILU. The output of the

local scale ensemble air quality forecasts will also be addressed using the same standards and methodologies. This task will ensure a standardised implementation of the uncertainty-enabled model. Web linkages, will implement service discovery tools within this use case and will interact directly with WP8 in defining and assessing these tools and links.

Active partners: NILU, AST

Legal Notices

The information in this document is subject to change without notice.

The Members of the UncertWeb Consortium make no warranty of any kind with regard to this document, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The Members of the UncertWeb Consortium shall not be held liable for errors contained herein or direct, indirect, special, incidental or consequential damages in connection with the furnishing, performance, or use of this material.

Executive Summary

This document brings together requirements from the application areas in which the UncertWeb solutions will be deployed and evaluated. In particular the following application domains are considered:

- Biodiversity and climate change (WP4);
- Land-use response to climatic and economic change (WP5);
- Short term uncertainty-enabled forecasts for local air quality (WP6);

The requirements for WP7 (D7.1) will follow later as a separate document, following the template used within this document.

These requirements will shape the design decisions made in WP1 (uncertainty representation), WP2 (chaining and discovery strategies) and WP8 (integration of UncertWeb components). The main aim is to establish the scope of models, input data and output data that will be considered within UncertWeb. Here by scope we mean a number of characteristics such as:

- Types of data being used (vector, gridded, observational, etc)
- Types of values being considered (continuous valued, categorical, binary, etc)
- Spatial and temporal support (domain and resolution) of the data
- How uncertainty is addressed, how important it is
- Physical size / structure of the data payload / time for the model computations

In addition we also explore the actual questions that will be addressed by the model chains constructed within UncertWeb and describe the analysis and post processing that might be employed in each chain. The results are presented in a standard format, describing the models and their inputs and outputs using prescribed tables.

The analysis of the requirements will be undertaken in subsequent deliverables; the aim of this document is to provide the requirements clearly. The requirements will inform the profiles of the information models and interface specifications that will be developed within UncertWeb.

Contents

1	Introduction	1
1.1	Definitions	1
2	Biodiversity application domain (WP4).....	2
2.1	Overview of the Model Chain	2
2.2	Components in the Model Chain	2
2.2.1	The e-Habitat model.....	2
2.2.2	Model components	5
2.2.3	Observation components	25
2.3	Questions the model chain would address.....	26
2.4	References	27
3	Land-use response to climatic and economic change (WP5).....	28
3.1	Overview of the Model Chain	28
3.2	Components in the Model Chain	29
3.2.1	The Climate Model.....	29
3.2.2	The Climate Data (observation component)	31
3.2.3	The Geophysical Data (observation component)	32
3.2.4	The Historical Crop Rotation data (observation component)	33
3.2.5	The Land Capability Classification System (Model Component).	38
3.2.6	The Markovian Crop Allocation Mechanism (model component).	45
3.2.7	The Economic Data (observation component).....	49
3.2.8	The Geometry of the Fields (observation component).....	51
3.2.9	The Field Use Simulator – LandSFACTS Model	52
3.2.10	Expected Yield Data (observation component)	63
3.2.11	The Yield Model	65
3.3	Questions the model chain would address.....	69
4	Probabilistic forecasting of air quality (WP6).....	71
4.1	Overview of the Model Chain	71
4.2	Components in the Model Chain	72
4.2.1	Model components	72
4.2.2	Observation components	83
4.3	Questions the model chain would address.....	85
5	Summary	86

1 Introduction

This report presents the consolidated requirements from the application domains for which we are developing prototypes within UncertWeb, namely:

- Biodiversity and climate change (WP4);
- Land-use response to climatic and economic change (WP5);
- Short term uncertainty-enabled forecasts for local air quality (WP6);

The work package on individual activity in the environment (WP7) will produce a separate requirements document which will be submitted separately. For each application domain a common format was used. First the context is given and the aims of the model chain are defined. Then the component models and data sources that are integrated within the chain are described along with key attributes needed to assist in defining the requirements for service chaining, tools to assist in service chaining and issues of data and uncertainty representation and propagation that will arise. These requirements will be extensively used within the project to define requirements in WP1, WP2 and WP3. They will also be made available to other projects.

Within this report there is no attempt to analyse the requirements in depth, this being the subject of later requirements documents to be delivered within WP1, WP2 and WP8.

1.1 Definitions

In order to ensure a standard set of terms is used throughout the following definitions are adhered to throughout the document.

Model web / Model chain: a chain of model components connected through web service interfaces.

Model component: a representation of a process or series of processes implemented as computer code, often called a simulator.

Model input: a value, or series of values (if a field), that must be provided to the model component to evaluate the model (likely to include parameters in the model component, initial and boundary conditions for the model component).

Model output: a value, or series of values (if a field), that is produced by the model when it is evaluated.

Observation component: a set of observations of reality that might be used as model inputs, but also might be used to compare against model outputs. The distinction between an observation component and a model input is that an observation component is used to describe observations (which are collected and stored – i.e there is some data there) which might be used in the processing chain somewhere, often for output validation. Model inputs are descriptions of what the models want to get as their inputs, which might not always be directly comparable to the observations we have available.

2 Biodiversity application domain (WP4)

This section reviews the requirements for the biodiversity application domain.

Frequently used acronyms (in this section)

AOO: Area of occupancy

APAAT: African Protected Areas Assessment Tool

EOO: Extent of Occurrence

GLWD: Global lakes and wetlands database

HRI: Habitat irReplaceability Index

HSI: Habitat Similarity Index

NDVI: Max NDVI – vegetation vigour, from near infrared and red reflectance

NDWI: Max NDWI - Presence of water, from near and short wave infrared reflectance

PA(s): Protected Area(s)

PoHS: Probability of Habitat Similarity

WDPA: World Database on Protected Areas

2.1 Overview of the Model Chain

The e-Habitat model is a pure statistical model, which in principle can take any kind of data. Although the original application (The Assessment of African Protected Areas - APAAT) was developed for a particular purpose, using 9 different variables as indicators of terrestrial habitats, the user of the planned e-Habitat model is free to choose other applications and other indicators. Even before an alpha-version of the system is available, new applications such as the analysis of the vulnerability of PAs to wild fires or the assessment of the representativity of marine monitoring stations has been suggested. The description of the model chain and the variables in this document will therefore be rather generic, although we also will include a description of the original 9 default variables used for the APAAT.

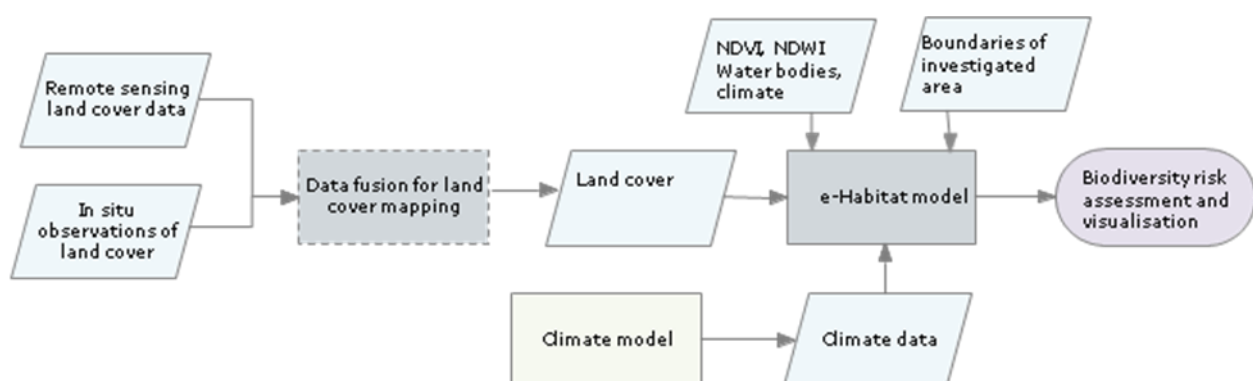


Figure 1. Flow chart for inputs, output and modelling steps in the e-Habitat model

2.2 Components in the Model Chain

2.2.1 The e-Habitat model

The computational part of the model chain in e-Habitat will only consist of one component computing a Probability of Habitat Similarity (PoHS) using as input variables the boundaries

of an area of reference and a set of thematic variables which will be used to model habitats inside and outside of this area. Figure 2 shows how, for a Protected Area (PA) located in Zambia (inset in the map of Africa), one can compute on a larger region the likelihood to find a similar set of combination of the selected variables and thus the likelihood to find habitats similar to the one of the PA.

Default thematic maps (1km raster):

- ✓ % tree cover
- ✓ % herbaceous cover
- ✓ % barren cover
- ✓ Elevation in metres
- ✓ Slope in degrees
- ✓ Aridity index
- ✓ % water body presence
- ✓ Normalized Difference Vegetation Index (NDVI)
- ✓ Normalized Difference Water Index (NDWI)

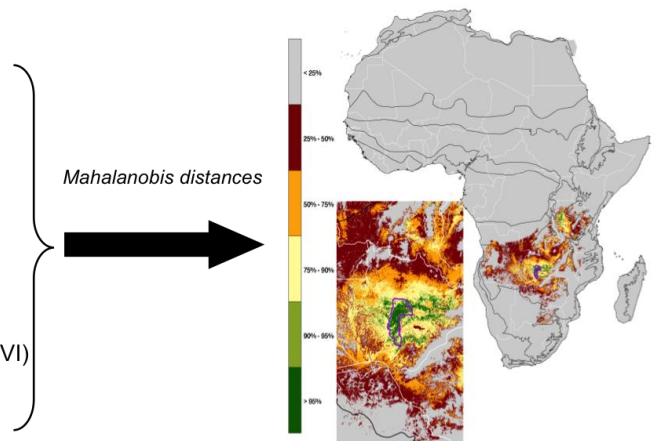


Figure 2. Flow Example of computation of the PoHS for the protected area of Kafue, Zambia –nine default themes are considered

The computation of the PoHS can be done by a large variety of methods, including Neural Networks, Support Vector Machine algorithms, Principal Components Analysis or Mahalanobis distances. Most frequently used in biodiversity related applications is the use of Mahalanobis distances. In ecological niche modelling, the PoHS is used in conjunction with species observations to compute a Habitat Similarity Index (HSI). The HSI yields for any location an index of its habitat similarity to the multivariate mean of habitat characteristics at locations where the analysed species was observed (Clark, *et al.*, 1993; Rotenberry, *et al.*, 2006). Hartley *et al.* (2007) used the PoHS to compute a Habitat irReplaceability index (HRI). The difference between the PoHS and the HRI is that the last requires an additional step. In the HRI, a pixel represents a potential replaceable habitat for the PA if the PoHS is significant ($p \geq 0.95$) and if the population density is less than or equal to the average population density with the PA. The total area¹ of the pixels outside the PA that meet these criteria is then summed and divided by the area of the PA to generate the HRI value. The smaller the HRI, the less replaceable a PA is, with a HRI of 0 suggesting that the PA habitat is unique and therefore irreplaceable. Conversely, a HRI greater than 1.0 suggests that there are potentially suitable habitats with a total area that is greater than the PA.

In Figure 2, there are only small patches outside the protected area that have high probabilities of being similar to the habitat inside the park. HRI will in this case be small, which indicates that the PA is highly irreplaceable and that it is important to maintain the park boundaries to assure the conservation of the habitats inside the park. Oppositely, if large areas outside the park could be considered similar to the habitats inside the park, it would be an indication of the habitats being able to survive without a strong protection of the PA.

The core of the HRI and HSI is the Mahalanobis distance **D** which is used as a measure of the similarity, and is defined as

¹ All data are projected to Mollweide Equal Area projection for the area calculations.

$$\mathbf{D} = \left(\sum_{v=1}^n \left([x_v - \mu_v]^T [\mathbf{C}]^{-1} [x_v - \mu_v] \right) \right)^{0.5}$$

where, \mathbf{x}_v is the value of the predictor variable \mathbf{v} , μ_v is the mean of variable \mathbf{v} for the protected area/habitat, and $[\mathbf{C}]$ is the covariance matrix for all \mathbf{n} variables in the PA. The covariance matrix for \mathbf{n} variables is given by

$$[\mathbf{C}] = \begin{bmatrix} \text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) & \cdots & \cdots & \text{cov}(x_1, x_n) \\ \text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) & \cdots & \cdots & \text{cov}(x_2, x_n) \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \text{cov}(x_n, x_1) & \text{cov}(x_n, x_2) & \cdots & & \text{cov}(x_n, x_n) \end{bmatrix}$$

and the covariance between any two variables, \mathbf{x}_1 and \mathbf{x}_2 , with means μ_1 and μ_2 and sample size \mathbf{m} is given by

$$\text{cov}(x_1, x_2) = \sum_{i=1}^m \left(\frac{(x_1 - \mu_1)(x_2 - \mu_2)}{m} \right)$$

The idea behind the Mahalanobis distance is that it gives the distance to a multivariate mean weighted by the covariance between the observations. Figure 3 is a simple example (taken from De Maesschalck et al. (2000))

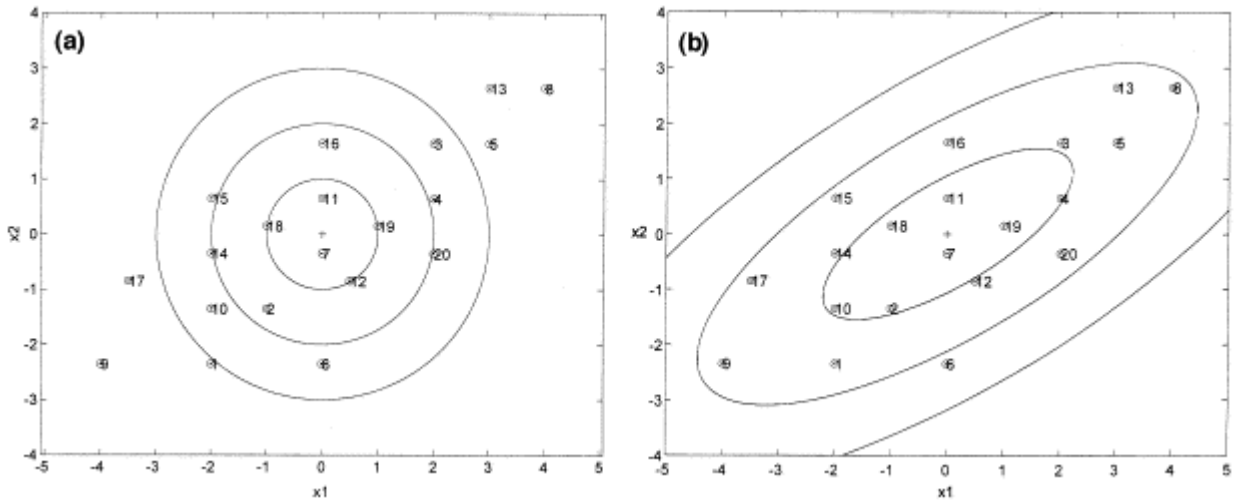


Fig. 3. (a) Plot of the simulated data for two variables x_1 and x_2 together with the circles representing equal Euclidian distances towards the center point. (b) Plot of the simulated data for two variables x_1 and x_2 together with the ellipses representing equal Mahalanobis distances towards the center point.

Note, that \mathbf{D} requires the *inverse* covariance matrix. When the predictor variables used to generate the mean vector and covariance matrix are normally distributed, then \mathbf{D} is distributed approximately according to a χ^2 distribution with $\mathbf{n}-1$ degrees of freedom, and so we can convert \mathbf{D} into \mathbf{p} -values. The \mathbf{p} -values (or probability values) range from 0.0 representing no similarity, through to 1.0 for areas which are identical. If the predictor variables are not normally distributed, then we can still make the conversion because it rescales the unbounded \mathbf{D} values to a 0.0 to 1.0 range.

This \mathbf{p} -value can be seen as the probability that a pixel outside the investigated area shares the same indicators as the one found for the selected area. A probability map showing the PoHS can so be returned to the user.

Because the Mahalanobis distance is an extremely simple measure of the similarity between locations outside and within a habitat/PA, it will be the first function implemented in e-Habitat to compute the PoHS. There are however a number of limitations of the method and we will so experiment with different versions of this component, all having similar input and outputs. Some data pre-processing might be necessary, and although not meant to be a part of the model chain, we need to address their uncertainties in the context of uncertWeb. Among the main improvements we will look at, we will provide means to

- analyse which indicators are most useful, possible automatic detection of these;
- weight the indicators according to their importance. This is likely to be application and maybe species specific.
- analyse more complex indicators allowing keeping temporal changes in view to assess ecological niche for migrating wildlife.
- address the limitations of the Mahalanobis distance for similarity assessments, i.e.
 - the assumption that the mean of the habitat best identifies the preferred habitat for a species
 - works with crisp habitat boundaries, whereas the population density of a species might be of large importance
- Explore other measures of similarity such as a stratified approach to the Mahalanobis distances (*Duncan and Dunn, 2001*), or the Maximum Entropy (*Jaynes, 1957; Phillips, et al., 2006*).

2.2.2 Model components

Models are described by a model table, and a number of input and output tables. These in turn might link to a number of relevant data component tables. Each model component will require several tables.

General information about the model component

Model name (make this unique)	e-Habitat (PoHS)
Model creator (where does it come from?)	Method derived from Clark et al.(1993), model implemented by Hartley et al. (2007)
Model licence (who can use it? Is it open source?)	Reimplementation
Model requirements (operating system, hardware)	R
Model computation (typical time to evaluate given a set of typical inputs – may require some commentary)	Very much depending on the resolution of the data set. The aim is to model 1*1 km grid, which will be almost 1 billion cells on a global scale. This will be computationally slow, but some problems may be solved through a more clever choice of resolution and/or tiling of the grids
Model description (<i>overview</i> of what the model does, how it works – this will need to be brief)	The model looks at the habitat of a species or a group of species (protected area). It is assumed that the species is found in the certain area because it is environmentally optimal. Using the Mahalanobis distance as a measure of similarity, the model estimates which other areas in the region that could possibly replace the habitat if it is threatened by climate change, deforestation, competition for land, population pressure, poaching etc.
Relevant references (papers or	Clark, J.D. , J.E. Dunn and K.G. Smith (1993), A

other resources, including web links, that describe the model)	<p>multivariate model of female black bear habitat use for a geographical information system, <i>Journal of Wildlife Management</i> 57, pp. 519–526</p> <p>Hartley, A., A. Nelson, P. Mayaux, and J. Grégoire, 2007. “The Assessment of African Protected Areas”. Luxembourg: Office for Official Publications of the European Communities, 2007. EUR 22780 EN</p> <p>See report and application at http://bioval.jrc.ec.europa.eu/PA/</p>
Website or link for downloading the model, if available.	The model & web service will be available in the future through http://dopa.jrc.ec.europa.eu/
Additional comments	

Data Sources

The e-Habitat model is, as a statistical model, not dependent on particular types of data input. It is rather a flexible model that can be used for very different purposes and with very different types of data. As the method has already been applied for African Protected Areas, and because we would like to test the method on a global scale and for different purposes, we will here describe a set of possible data sets. Despite not being exhaustive, this list should give a general overview of the sources and types of data, and their uncertainties, which will be of use also when different data sets are used. The numbers in the column for input name refer to the number of the tables below. Some data sets are currently not used, but are still included in the summary, as they might be included in the model at a later stage. Only the landcover data set is described in detail in a separate table (table 8).

Summary of component model inputs. The importance rank is 1- more important, 10 least important.

Input name	Short description	Uncertain?	Importance rank
Country borders	VMap0 data library - The CIAMAP database and National Geospatial-Intelligence Agency (<i>NGA</i>) former NIMA (-> 2003)	Yes	10
(1) Protected areas	World Database on Protected Areas (WDPA), managed by UNEP-WCMC	Yes – location and existence	2
Eco-regions	Vegetation map of Africa (UNESCO)	Yes – location	10
(9) Mammals (280)	Extent of occurrence –African mammals data bank, Boitani et al 1999 (EC funded)	Yes – location	3
Amphibians (all 5918)	Extent of occurrence –IUCN Global Amphibian Assessment, IUCN et al. 2006	Yes – location	3
Birds (1591 threatened)	Extent of occurrence –BirdLife International’s Important Bird Areas	Yes – location	3

	(IBAs) - Also checked against 2006 IUCN Red List of Threatened Species		
(2) Vegetation cover (for wood, herbaceous and bare ground)	Derived from all seven bands of MODerate-resolution Imaging Spectroradiometer	Yes	3
(3) Elevation	Shuttle Radar Topography Mission (SRTM 30)	Yes	3
(4) Slope	Shuttle Radar Topography Mission (SRTM 30)	Yes	3
(5) Climate (aridity)	Calculated from WorldClim data as $I_h = R/E$ where R is annual precipitation and E is annual potential evapotranspiration. Global climate grids (1 km) – interpolated from Global Historical Climate Network Data set, WMO climatological normals (CLIMO), FAOCLIM, International Center for Tropical Agriculture (CIAT), additional local data sources	Yes	3
(6) Vegetation vigour	Max NDVI (calc from near infrared and red reflectance) – satellite images	Yes	3
(6) Presence (absence) of water	Max NDWI (calc from near infrared and short wave infrared reflectance) – satellite images	Yes	3
(7) Water bodies	Global lakes and wetlands database (GLWD) – different sources	Yes – location?	3
(7) Small water bodies – apx 1 km ² – percentage of time SWBD (SRTM Water body data?)	SPOT-VEGETATION images? In case of disagreement, GLWD is prioritized over SWBD	Yes – location and percentage	3
(8) Land cover (27 categories)	Global Land Cover 2000 (Mayaux et al. 2004) - Obs from 2000 from SPOT-4 satellite	Yes – categorical	3
Population density	Gridded population of the world GPWv3	Yes –	4
Roads	Vmap0 (NGA)	Yes –	4
Urban areas and populated places	Global Rural-Urban Mapping Project (GRUMP)	Yes –	4

We give most indicators the same importance because we do not yet have a method for weighting the indicators; they all have the same weight in the model, independent on their real impact on the animals' habitats. Similarly, e-habitat can be perfectly used without any information on species distributions if one is interested only in forecasting changes to a specific habitat (e.g. rainforest) because of anthropogenic pressure.

The default set of variables described above is used by the APAAT and we would need to have similar layers at the global scale to do habitat modelling beyond the African continent. This possibility will be considered in a second stage and the availability of similar data

investigated. As a proof of concept, we will in a first stage analyze protected areas and habitats considering only Holdridge's lifezones (*Holdridge, 1947*). Holdridge suggested that it was sufficient to use three variables for classifying land areas:

Mean annual biotemperature (logarithmic)	Can be derived from WorldClim monthly temperature grids	Yes – updated?	4
Annual precipitation (logarithmic)	From WorldClim	Yes – updated?	4
Ratio of annual Potential evapotranspiration to mean annual precipitation	Can be derived from monthly temperature data from WorldClim	Yes – updated?	4

A schema showing the classification scheme is shown below

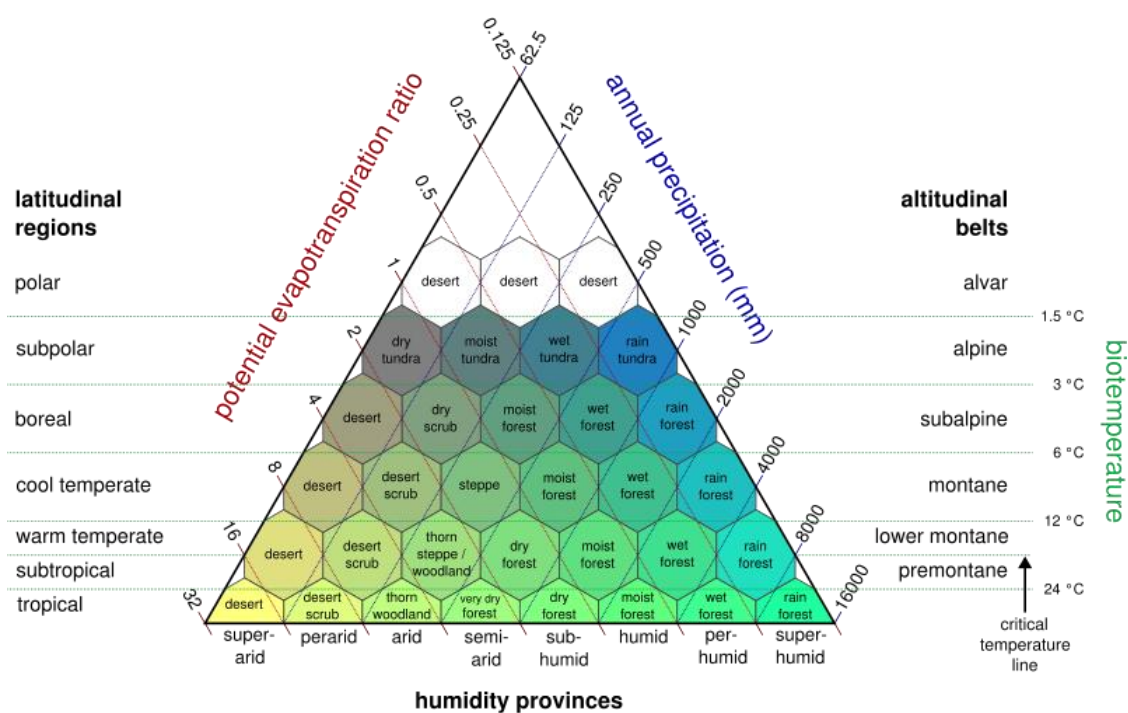


Figure 3. Holdridge life zone classification scheme. Although conceived as three-dimensional by its originator, is usually shown as a two-dimensional array of hexagons in a triangular frame. Figure from Wikipedia, http://en.wikipedia.org/wiki/Holdridge_life_zones

The mean biotemperature is based on the growing season length and temperature. It refers to all temperatures above freezing, with all temperatures below freezing adjusted to 0 °C, as plants are dormant at these temperatures. All of these variables can be derived from WorldClim data, which are described in Table 5.

Possible variables not included in the list above

- Variability of habitat (species in need of a changing environment – or with different requirements at different times of year) - migrating species
- In the case e-Habitat is used for ecological niche modelling, competition between species will have to be considered
- The habitat might not be optimal, the species just didn't find a better place

- Other derived variables, such as distance to water features. These require an excellent data set on rivers and water bodies and fast calculation of travel times from points to the closest water body/river. The feasibility of generating these datasets within the framework of this project is questionable.

Description of uncertain model inputs (*note one table per input that we will treat as uncertain in our model web, and thus included in the uncertainty propagation*).

Table 1

Input name (unique if possible – add model name e.g. model:input)	e-habitat:SelectedArea
Input role (parameter in model, initial condition, boundary condition)	Used to define the area for which the indicators are assumed to be optimal
Input type (continuous, categorical, binary etc)	Categorical (vector data or raster with inside vs. outside)
For spatial inputs what is the domain (extent of the input region)	Global
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	Usually vector-data
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Size of area
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	NA
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	NA
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	NA
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	Currently unknown as the selected area of interested can be defined in any way. For reporting purposes, boundaries of PAs would be used as input data and uncertainty could refer to the encoding of PA boundaries and to the existence of PA. Uncertainty can also come from the subdivision of PAs into different sets with different legal status
Is the marginal (at a single location or object) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No This is a possible classification error, maybe accessible through expert elicitation
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No, and we will not try to estimate this pdf
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	No
Do you know, and can you specify typical ranges for this input?	0.0001 – 10.000.000 km ²
Input source (describe where this input might come from – link to a observation component if this makes sense)	World Data base on Protected Areas -UNEP WCMC - http://www.wdpa.org/Download.aspx

Table 2

Input name (unique if possible – add model name e.g. model:input)	e-Habitat:pTreeCover (similar for percentage herbaceous vegetation and bare ground)
Input role (parameter in model, initial condition, boundary condition)	Used as one of the indicators for describing the habitat
Input type (continuous, categorical, binary etc)	Continuous
For spatial inputs what is the domain (extent of the input region)	Global
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	30 arc-seconds (apx 1*1 km grid close to equator, usually referred to as a 1 km ² grid)
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Grid cell mean
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	NA
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	NA
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	~1 year average (Oct 31, 2000 – Dec 9, 2001)
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	Uncertainty is partly given, in the form of standard error (in %) as a function of regression tree nodes (only tree cover ~ 12 %), and as R ² from different validations (0.81-0.94, dependent on scale)
Is the marginal (at a single location or object) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No, but we might be able to approximate the distribution from the R ² values above
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No
Do you know, and can you specify typical ranges for this input?	0-100 (percentage of tree cover)
Input source (describe where this input might come from – link to a observation component if this makes sense)	Derived from Modis images: http://glcf.umiacs.umd.edu/data/vcf/

Table 3

Input name (unique if possible – add model name e.g. model:input)	e-Habitat:elevation
Input role (parameter in model, initial condition, boundary condition)	Used as one of the indicators for describing the habitat
Input type (continuous, categorical, binary etc)	Continuous
For spatial inputs what is the domain (extent of the input region)	Global
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	Down to 30 m grid resolution
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Mean elevation
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	NA
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	NA
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	NA
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	Uncertainty has been estimated on the basis of validation points. It is for the 30 m version in the order of a few meters for all continents, std is 3.5-5.9 meters. The report suggest that the errors are on the optimistic side though, as very rugged terrain is likely to have been undersampled for the validation
Is the marginal (at a single location or object) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	We can probably assume normal distribution
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	Some work was done to assess the spatial correlation of the errors, empirical structure functions and correlation functions for all continents were derived.
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	Being a high-resolution space filling data set, empirical pdfs are possible to derive. Importance is most likely limited
Do you know, and can you specify typical ranges for this input?	-300 - 8900 m
Input source (describe where this input might come from – link to a observation component if this makes sense)	Shuttle Radar Topography Mission (NASA) http://www2.jpl.nasa.gov/srtm/

Table 4

Input name (unique if possible – add model name e.g. model:input)	e-habitat:slope
Input role (parameter in model, initial condition, boundary condition)	Used as one of the indicators for describing the habitat
Input type (continuous, categorical, binary etc)	Continuous
For spatial inputs what is the domain (extent of the input region)	Global
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	Down to 30 m grid resolution
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Same as resolution. Slope can be found in different ways between neighbouring grid cells
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	NA
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	NA
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	NA
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	NA
Is the marginal (at a single location or object) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No. This is also a value that is extremely scale dependent, decreasing with increasing grid cell size
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No, although it might be possible to derive it from the elevation grid through error propagation. Questionable whether it is worth the effort.
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No
Do you know, and can you specify typical ranges for this input?	0-90 degrees, although majority will be closer to 0 and 90 is only asymptotically possible
Input source (describe where this input might come from – link to a observation component if this makes sense)	Derived from Shuttle Radar Topography Mission (NASA) http://www2.jpl.nasa.gov/srtm/

Table 5

Input name (unique if possible – add model name e.g. model:input)	e-Habitat:climate
Input role (parameter in model, initial condition, boundary condition)	Used as one of the indicators for describing the habitat, annual rainfall divided by annual potential evapotranspiration
Input type (continuous, categorical, binary etc)	Continuous
For spatial inputs what is the domain (extent of the input region)	Global
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	30 arc-seconds (apx 1*1 km grid close to equator, usually referred to as a 1 km ² grid)
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Grid cell mean
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	Averages based on the period 1950-2000
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	Mostly monthly averages
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	Monthly (annual for some)
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	The WorldClim data set was interpolated with kriging. The calculation of climate (aridity) from the data set is non-linear, hence is the derivation of uncertainty for the climate index not straightforward. Simulations might be an option if we get access to original data
Is the marginal (at a single location or object) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	To some degree, see above, from interpolation error or simulations
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	A correlation function has been estimated from point data
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	No
Do you know, and can you specify typical ranges for this input?	The index ranges from 0 (hyper-arid < 0.03) to more than 1 (Extremely humid > 1)
Input source (describe where this input might come from – link to a observation component if this makes sense)	WorldClim.org

Table 6

Input name (unique if possible – add model name e.g. model:input)	e-Habitat:NDVI (and NDWI)
Input role (parameter in model, initial condition, boundary condition)	Used as one of the indicators for describing the habitat
Input type (continuous, categorical, binary etc)	Continuous
For spatial inputs what is the domain (extent of the input region)	Global
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	30 arc-seconds
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	30 arc-seconds
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	1998-2005 (for the test case)
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	NA
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	10 days
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	We can choose to see this as a certain input, as it is just an index. The uncertainty of how this is related to the vegetation vigour is of course larger and unknown
Is the marginal (at a single location or object) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	No
Do you know, and can you specify typical ranges for this input?	0-255
Input source (describe where this input might come from – link to a observation component if this makes sense)	NASA: http://neo.sci.gsfc.nasa.gov/Search.html?datasetId=MOD13A2_M_NDVI

Table 7

Input name (unique if possible – add model name e.g. model:input)	e-Habitat:waterBodies
Input role (parameter in model, initial condition, boundary condition)	Used as one of the indicators for describing the habitat
Input type (continuous, categorical, binary etc)	Continuous (percentage)
For spatial inputs what is the domain (extent of the input region)	30 arc-seconds
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	30 arc-seconds
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Grid cell mean
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	2000-2005 (for example case)
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	NA
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	Annual average
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	No
Is the marginal (at a single location or object) probability density function known for this input?	No
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	No
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No
Do you know, and can you specify typical ranges for this input?	0-1 (percentage of water bodies within pixel)
Input source (describe where this input might come from – link to a observation component if this makes sense)	Global Lakes and Wetlands Database (GLWD) http://www.worldwildlife.org/science/data/item1877.html Data generated from VEGETATION satellite images, described in EU Report 22344

Table 8

Input name (unique if possible – add model name e.g. model:input)	e-Habitat:landcover (GLC2000 map)
Input role (parameter in model, initial condition, boundary condition)	Used as one of the indicators for describing the habitat
Input type (continuous, categorical, binary etc)	Categorical
For spatial inputs what is the domain (extent of the input region)	Global
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	(most likely) 30 arc-seconds
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Grid cell average converted to land cover class
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	14 months (2000)
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	NA
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	NA
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	From validation tests (Northern Eurasia only) – as percentage of wrongly classified pixels. Depending on land cover class, from 0 percent (water, recent burns) to ~ 20 % (Light everg, Dark everg and decid needle forest) and 42% for Confireous shrubs
Is the marginal (at a single location or object) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No
Do you know, and can you specify typical ranges for this input?	27 different classes, only partly ordered
Input source (describe where this input might come from – link to a observation component if this makes sense)	Global Land Cover 2000 http://bioval.jrc.ec.europa.eu/products/glc2000/products.php

Using e-Habitat for ecological niche modelling

One of the largest potentials of e-Habitat comes from its possible use for ecological modelling. Table 9 describes the species data that would be used.

Table 9

Input name (unique if possible – add model name e.g. model:input)	e-Habitat:EEO-maps (Extent of Occurance) for mammals
Input role (parameter in model, initial condition, boundary condition)	Used to define the area for which the indicators are assumed to be optimal
Input type (continuous, categorical, binary etc)	Categorical – animal exists yes/no
For spatial inputs what is the domain (extent of the input region)	Global
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	Vector data or 30 arc-seconds raster data
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Existence of animal within grid cell
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	NA
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	NA
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	NA
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	Currently not. Uncertainty depends on existence of limiting feature (rivers, mountains), on possibility to observe animal (small in jungle versus big in open landscape), and on method to infer the boundaries from observations. See below.
Is the marginal (at a single location or object) probability density function known for this input?	No
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	No
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	No
Do you know, and can you specify typical ranges for this input?	0.001 km ² – 25.000.000 km ²
Input source (describe where this input might come from – link to a observation component if this makes sense)	African mammals data bank (University of Rome) and IUCN Redlist Global data: http://www.gisbau.uniroma1.it/amd/framed_ata.html http://www.iucnredlist.org/technical-documents/spatial-data

A more thorough description of this input is given here, as the Extent of Occurrence maps will be of importance both as an input and during a possible validation. A proper assessment of uncertainty is, however, rather difficult, as is the consideration whether to use derived maps of Area of Occupancy with better description of uncertainties, but also with a risk of spurious correlation in the validation process.

There are two types of maps that can be used for describing the habitats of animals. The first are the Extent of Occurrence maps, which gives the borders of known observations of the species. The extent of occurrence maps will be of very different quality, depending on the species. The original sources of the maps are often difficult to identify. An extensive job was done to set up the African mammals data bank (*Boitani, et al., 1999*), and to validate the extent of occurrence maps. Most of the maps were derived from different “authoritative” sources, sometimes in the form of a scientific paper or an IUCN reports but often the maps come from safari field guides, and have been revised by experts. The origin is often difficult or impossible to find. The following is the source description of the EOO for *Phacochoerus aethiopicus* (Somali or desert warthog) from Oliver (1993):

“This geographic representative of the Cape warthog is recorded from Somalia, both in the north and in Jubaland in the south, and from northern Kenya. Both this species and the common Warthog have been obtained in northern Somalia, where locality records for the common species form an enclave in the vicinity of Berbera, with sparse records of Somali warthog to the west, east and south. The two species may be parapatric or even partly sympatric and ecologically segregated in northern Somalia, but this has yet to be confirmed. Their relative geographical disposition in Kenya (or eastern Ethiopia) cannot be assessed at all in the absence of adequate specimens or information.”

The maps for this species were then revised by an expert, Dr. J.P. D’Huart. No further references are available, as for many other species. Another example is Hippopotamus amphibious (common hippopotamus), where the same report Oliver (1993) cite some old and difficult to obtain sources, such as Kock (1970). It is interesting to notice that the occurrence map for the common hippopotamus only include areas close to major waterways (which leads to a very patched habitat), whereas habitats for most other animals appear more connected, although similar distributions are probable.

The project for the African Mammals Databank have also set up a South-Asian Mammals Databank, and is currently busy with a Global Mammals Assessment. This will be available in a few years. In this databank, the EOO maps will be derived on the basis of a consensus of experts. Uncertainty will most likely not be a part of the assessment. As these maps will be used together with land cover maps to produce (the for many purposes more useful) Area of Occupancy maps (AOO) (see below), it seems more likely that the borders are drawn on the large side.

The AOO-maps are derived on the basis of EOO-maps and the assumed suitability for the species. Whereas the EOO-maps are large polygons that might include areas not suitable for the species (such as densely populated areas or environmentally not suitable areas), the AOO-maps should only include areas where it is possible to encounter the species. Below is a description of the two types of maps from the 2001 IUCN Red List Categories and Criteria version 3.1

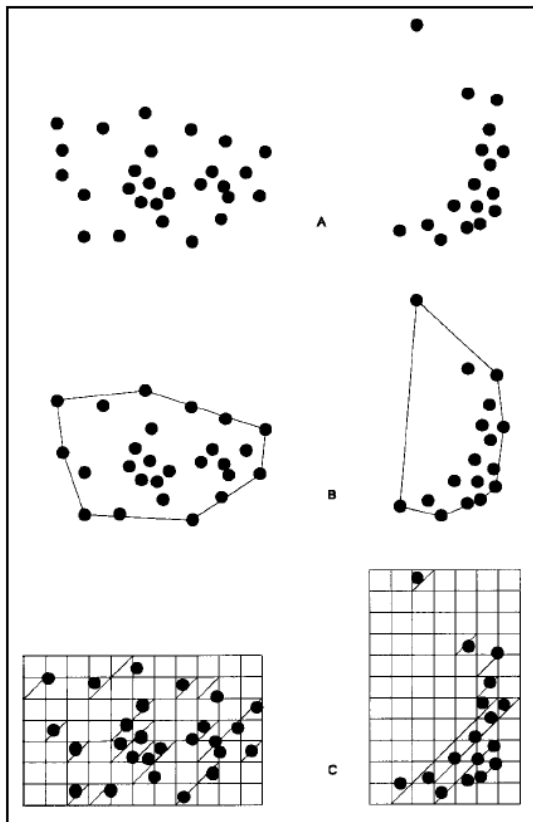


Figure 2. Two examples of the distinction between extent of occurrence and area of occupancy. (A) is the spatial distribution of known, inferred or projected sites of present occurrence. (B) shows one possible boundary to the extent of occurrence, which is the measured area within this boundary. (C) shows one measure of area of occupancy which can be achieved by the sum of the occupied grid squares.

“Extent of occurrence (Criteria A and B)

Extent of occurrence is defined as the area contained within the shortest continuous imaginary boundary which can be drawn to encompass all the known, inferred or projected sites of present occurrence of a taxon, excluding cases of vagrancy (see Figure 2). This measure may exclude discontinuities or disjunctions within the overall distributions of taxa (e.g. large areas of obviously unsuitable habitat) (but see 'area of occupancy' below). Extent of occurrence can often be measured by a minimum convex polygon (the smallest polygon in which no internal angle exceeds 180 degrees and which contains all the sites of occurrence).

Area of occupancy (Criteria A, B and D)

Area of occupancy is defined as the area within its 'extent of occurrence' (see previous section) which is occupied by a taxon, excluding cases of vagrancy. The measure reflects the fact that a taxon will not usually occur throughout the area of its extent of occurrence, which may contain unsuitable or unoccupied habitats. In some cases (e.g. irreplaceable colonial nesting sites, crucial feeding sites for migratory taxa) the area of occupancy is the smallest area essential at any stage to the survival of existing populations of a taxon. The size of the area of occupancy will be a function of the scale at which it is measured, and should be at a scale appropriate to relevant biological aspects of the taxon, the nature of threats and the available data. To avoid inconsistencies and bias in assessments caused by estimating area of occupancy at different scales, it may be necessary to standardize estimates by applying a scale-correction factor. It is difficult to give strict guidance on how standardization should be done because different types of taxa have different scale-area relationships.”

When validating the AOO-maps of the African Mammals Databank (Boitani, et al., 1999), the authors picked four countries where the data were validated against field observations done

by locally collaborating institutes. Altogether 427 plots were chosen for validation, and validation analysis was done for those species for which at least 1% of the known Extent of Occurrence is included in the sample areas. This will of course not cover all species, but the authors made an attempt to cover different environmentally diverse regions, although it was also deemed necessary to choose countries where it was possible to find local collaborators and good quality data on natural resources. The chosen countries were Botswana, Cameroon, Uganda and Morocco, and 181 species were available for validation within these countries.

The measure used for validation is the Index of Accordance, which was the percentage of correctly predicted plots (both presences and absences) within the Extent of Occurrence. This index of accordance varies between species within each order. The average is around 60%, which still means that a large amount of the plots are misclassified.

For our purposes, it could be more relevant to use the AOO-maps both for input and output (validation), as these maps better describe the real habitats of the species. The risk is that we “reuse” some of the information used to create the AOO-maps in our replaceability mapping. This is acceptable for the mapping purposes, but could easily lead to spurious correlation for the validation case, indicating a much better model performance than in reality, as the habitat model would not be independent from the method used to produce the AOO-maps.

Summary of component model outputs

Output name	Short description	Importance rank
PoHS	The probability of habitat similarity shows, for an area of reference, the probabilities to find areas elsewhere with similar composition of indicators	1
HRI	Habitat iRreplaceability index – the sum of pixels representing a suitable replaceable habitat, divided by original habitat size	2

Description of uncertain model output

Output name in form model:output	e-Habitat:PoHS
Output role (feeds into / is used for)	The Probability of Habitat Similarity shows, for an area of reference, the probabilities to find areas elsewhere with similar composition of indicators
Output type (continuous, categorical, binary etc)	Continuous (Probability 0-1)
For spatial outputs what is the domain (extent of the input region)	Global
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	Same as for input raster – planned is 30 arc-seconds
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Same as for input raster – planned is 30 arc-seconds
For temporal outputs what is the domain (extent of the input time – at a point, over a time interval?)	NA
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	NA
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	NA
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble, distribution)?	No, not currently known. This is one of the targets of UncertWeb
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	It is not known at the moment, but will probably be analysed during the time of the project
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support.	There is currently no validation method for the result
Will visualisation be necessary? How might this be done?	Visualisation will be necessary, as maps
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	No, it is currently not cross-validated. Cross-validation might be possible for assessment of species habitats, less for protected areas

Output name in form model:output	e-Habitat:HRI
Output role (feeds into / is used for)	Habitat replaceability Index – used to assess the size of possible replacement area
Output type (continuous, categorical, binary etc)	Continuous (0-Inf)
For spatial outputs what is the domain (extent of the input region)	NA
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	NA
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	NA
For temporal outputs what is the domain (extent of the input time – at a point, over a time interval?)	NA
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	NA
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	NA
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble, distribution)?	No, not currently known. This is one of the targets of UncertWeb
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	It is not known at the moment, but will probably be analysed during the time of the project
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support.	There is currently no validation method for the result
Will visualisation be necessary? How might this be done?	Visualisation will be necessary, as maps
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	No, it is currently not cross-validated. Cross-validation might be possible for assessment of species habitats, less for protected areas

2.2.3 Observation components

These are the observations that might be used to determine some inputs within UncertWeb, and in particular for validation of outputs. Each application WP is likely to need several of these, and some are likely to be shared. They are separate from model components since they might exist as services outside the particular model chain.

Data set name (unique)	e-Habitat:EEO-maps (Extent of Occurrence) (See also table 9 for inputs)
Data source (where does the data come from)	African mammals data bank (University of Rome) and IUCN Red list Global data:
Data availability (where can this be obtained from, access rights)	From Web (see below). From terms of use for IUCN data: “You are hereby granted a license to use, download and print the materials contained in The Red List solely for conservation purposes, scientific analyses, or research provided that such use is in accordance with this User Agreement” AMD does not give terms of use
Data accessibility (is this available over the web, and how is this made available?)	http://www.gisbau.uniroma1.it/amd/framedata.html http://www.iucnredlist.org/technical-documents/spatial-data
Data type (continuous, categorical, binary etc)	Categorical – animal exists yes/no
For spatial observations what is the domain (extent of the input region)	Global
For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)	Vector data or 30 arc-seconds raster data
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Existence of animal within grid cell
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	NA
For temporal observations what is the sampling frequency.	NA
For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	NA
Is uncertainty known and is this broken down into observation errors, processing errors, representativity?	No
Do you have confidence in being able to specify uncertainties? What issues can you foresee in specifying uncertainties?	No, see also last paragraph on AOO and EOO in input section above.

2.3 Questions the model chain would address

The idea behind e-Habitat is mainly to identify other regions that have similar indicators (climate, terrain, water resources) as we can find within a habitat or protected area (PA) (Hartley, *et al.*, 2007). This is done by estimating the similarity between points outside the habitat and locations inside the habitat. The aim is to identify habitats or PAs that (in theory) could be replaced by similar areas elsewhere, and habitats and PAs that are unique. The use of this can be different:

- For habitats of a species, identification of a habitat elsewhere, which is not used by the animal, can be used in analysis of the requirements and competitive environment of the animal
- For habitats of an animal, similar areas elsewhere (with a lower density of the animal) could be used for relocation of animals in case of threats by population pressure etc. The possibilities and consequences for this will of course have to be examined in more detail than this model is able to do.
- For PAs, identification of similar areas can suggest extension of the protection zone, or limit the need for funding of this park.
- For PAs, oppositely, the identification of a PA as unique can be an incentive for further funding and enhancement of the protection of this area

The case study will be an analysis of the replaceability for habitats and protected areas in Africa. The original application was a “one-time-assessment only” that we want to automate in view to get:

- The possibility to change the indicators in a simple manner
- The possibility to use the method for analysing the impact of each indicators (leave-one out cross-validation approach to habitat modelling)
- The possibility to simply being able to reuse the method for other purposes
- The possibility to estimate uncertainty in addition to the variable itself
- The possibility to have an ensemble approach to habitat modelling

The ability to estimate uncertainty is of course a key issue. One challenge in the uncertainty analysis might be that we only know to some degree the uncertainty of the input raster maps. Also for the inputs where we do know the uncertainty, we might be in the need of simulations of a type that should preferably be made on the basis of the original data. The most easily available should be the climate data, where we have already asked the responsible person whether he would be willing to help us, either by creating the simulations for us, or by making the original data and methods available.

2.4 References

- Boitani, L., F. Corsi, A. De Biase, I. D'Inzillo Carranza, M. Ravegli, G. Reggiani, I. Sinibaldi, and P. Trapanese (1999), A databank for the conservation and management of the African mammals, 1155 pp, Institute of Applied Ecology and European Commission, DG for Development, Rome.
- Clark, J. D., J. E. Dunn, and K. G. Smith (1993), A multivariate model of female black bear habitat use for a geographical information system, *Journal of Wildlife Management* 57, 519-526.
- Duncan, L., and J. E. Dunn (2001), Partitioned Mahalanobis D2 to improve GIS classification, in *Proceedings of the SAS Users Group International Number 26*, edited, pp. Paper 198-126, SAS Institute, Cary, North Carolina, USA.
- Hartley, A. J., A. Nelson, P. Mayaux, and J.-M. Grégoire (2007), The assessment of African protected areas, JRC Scientific and Technical Reports, EUR 21296, 70 pp, Office for Official Publications of the European communities, Luxembourg.
- Holdridge, L. R. (1947), Determination of world plant formations from simple climatic data, *Science*, 105, 367-368.
- Jaynes, E. T. (1957), Information theory and statistical mechanics, *Phys. Rev.*, 106, 620-630.
- Kock, D. (1970), Die Verbreitungsgeschichte des Flusspferdes, *Hippopotamus amphibius* Linne, 1758, im unteren Nilgebiet, *Saugtierk. Mitt*, 18, 12-25.
- De Maesschalck, R., D. Jouan-Rimbaud, D. L. Massart (2000) The Mahalanobis distance *Chemometrics and Intelligent Laboratory Systems*, 50 (1), pp. 1-18
- Oliver, W. L. R. (Ed.) (1993), Pigs, peccaries, and hippos. Status survey and conservation action plan., IUCN/SSC Pigs and Peccaries Specialist Group & IUCN/SSC Hippo Specialist Group
- Phillips, S. J., R. P. Anderson, and R. E. Schapire (2006), Maximum entropy modelling of species geographic distributions, *Ecological modelling*, 190, 231-259.
- Rotenberry, J. T., K. L. Preston, and S. T. Knick (2006), GIS-based niche modelling for mapping species' habitat, *Ecology*, 87, 1458-1464.

3 Land-use response to climatic and economic change (WP5)

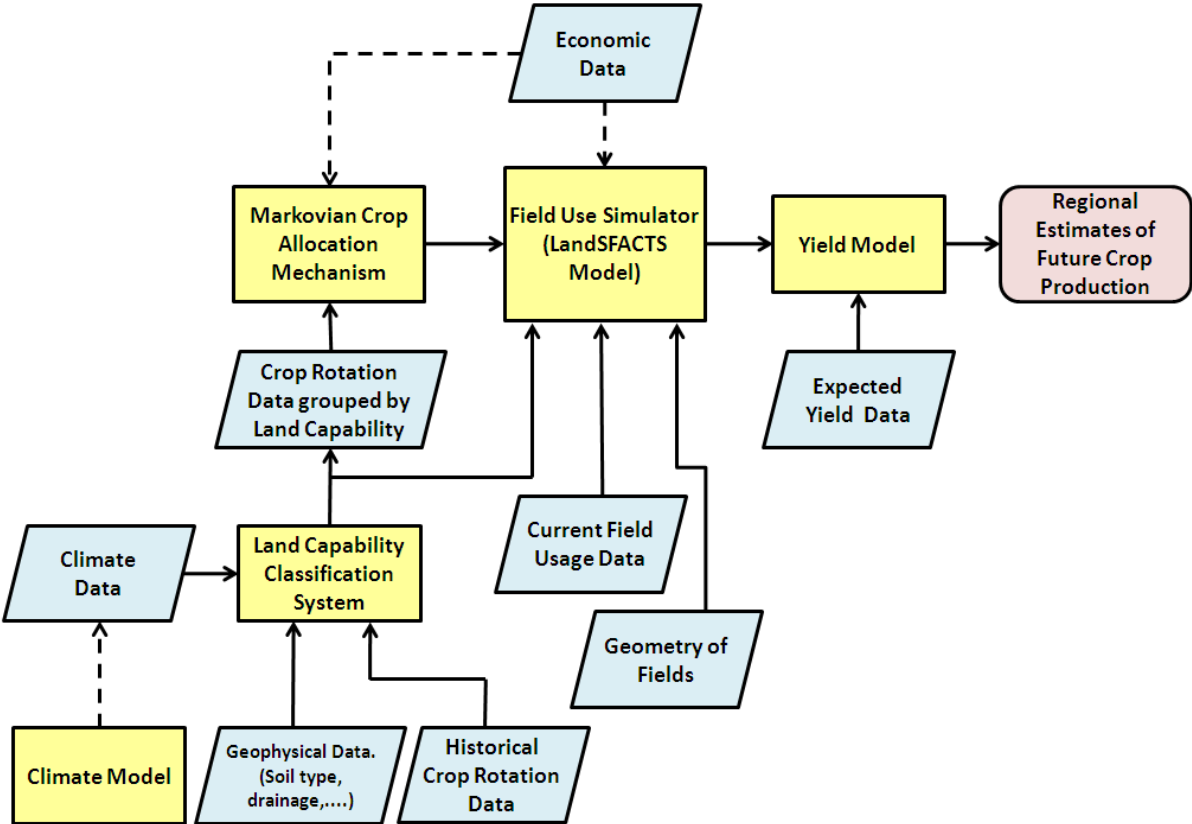
This section reviews the requirements for the land-use response application domain.

Frequently used acronyms (in this section)

- CAM: Crop Allocation Mechanism
- Defra: Department for the Environment, Food and Rural Affairs
- IACS: Integrated Administration and Control System
- JCA: Joint Character Areas
- LandSFACTS: Landscape Scale Functional Allocation of Crops Temporally and Spatially
- LCCS: Land Capability Classification System
- RLR: Rural Land Register
- UKCP09: United Kingdom Climate Projections 2009

3.1 Overview of the Model Chain

The following model chain diagram represents the model chain for the land use case study in WP5:



Firstly, we assess the land capability through the **Land Capability Classification System**, which takes climate data (either current or some simulated future climate scenario gained from the climate model), geophysical data (such as soil type and field drainage) and the historical crop rotation data for the fields in the region under study to determine which types

of crops can be grown in each field. This system groups the historical crop rotation data by land capability and outputs the data in a grouped form, as well as returning a listing of the land capability classification determined for each field parcel.

This historical crop rotation data grouped by land capability, and possibly economic data, then feeds into the **Crop Allocation Mechanism** to compute transition matrices for crop rotation for the fields in the region of study under each land capability.

The transition matrices, the geometry of the fields in the region of study, the land capability for each of these fields, data on current field usage for the fields and possibly economic data and economic constraints are then all used as inputs to the **Field Use Simulator**. Here, we are using the **LandSFACTS Model** as our simulator. This model produces an overall simulated land-use scenario for the region under study, which is then fed into a **Yield Model** along with data on expected crop yields to produce the overall output of the model chain: regional estimates of future crop production.

3.2 Components in the Model Chain

Component Ordering:

The components of our model chain are given in the following order:

- The Climate Model and the Climate Data.
- The Geophysical Data, the Historical Crop Rotation Data (which includes the Current Field Usage data as the final year of this data)
- The Land Capability Classification System and the output from this – the Crop Rotation Data grouped by Land Capability.
- The Markovian Crop Allocation Mechanism.
- The Economic Data and the Geometry of Fields.
- The Field Use Simulator – LandSFACTS Model.
- The Expected Yield Data and the Yield Model, including the final output of the model chain: The Regional estimates of Future Crop Production.

3.2.1 The Climate Model

We intend to use data from the UKCP09 Climate model. We will not run the climate model itself, but we will use projections generated for this model. Therefore, we have used the observation component table to describe this.

Data set name (unique)	UKCP09 projections
Data source (where does the data come from)	These data are the product of a statistical analysis from the Met Office’s regional climate model.
Data availability (where can this be obtained from, access rights)	This source of data is free to all researchers after submission of a form regarding the use of the data.
Data accessibility (is this	This data is provided through a web service at

available over the web, and how is this made available?)	http://ukclimateprojections.defra.gov.uk/
Data type (continuous, categorical, binary etc)	The data are continuous as they are measures of climatic phenomenon (e.g. temperature, wind speed, precipitation).
For spatial observations what is the domain (extent of the input region)	The data we are interested in cover all of the UK.
For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)	The data can be accessed in three different spatial formats: 25km grid squares, administrative regions (16 for the UK) or river basins (23 for the UK). The latter two are averages over the 25km grid squares that fall under the regions. There are complete sets of data for each location.
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	See previous.
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	N/A
For temporal observations what is the sampling frequency.	N/A
For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	The UKCP09 projections can produce annual, seasonal, monthly and daily predictions (although the use of daily predictions is not recommended).
Is uncertainty known and is this broken down into observation errors, processing errors, representativity? How is it described?	The uncertainty is demonstrated through an ensemble of runs that have been generated using emulation techniques. The variation in the ensemble is caused by perturbations of the inputs to the climate model. There is not any formal account of modelling uncertainty (either stemming from the emulator or from the model-reality discrepancy).
Do you have confidence in being able to specify uncertainties? What issues can you foresee in specifying uncertainties?	We do not think there is scope to improve the characterisation of uncertainty in these data.

3.2.2 The Climate Data (observation component)

Data set name (unique)	Current temperature and rainfall data (CurWeather)
Data source (where does the data come from)	UK Meteorological Office. MIDAS Land Surface Stations data
Data availability (where can this be obtained from, access rights)	The Met Office wish to monitor the use of this data and require an acknowledgement of the data source if they are used in any publication. Their data is therefore restricted and, as for UKCP09, user must register prior to gaining access.
Data accessibility (is this available over the web, and how is this made available?)	This data is provided through a web service at http://badc.nerc.ac.uk/data/ukmo-midas/
Data type (continuous, categorical, binary etc)	Continuous – average rainfall and temperatures for certain periods of the year. (However, we have also considered start and end of growing season, which has different properties.)
For spatial observations what is the domain (extent of the input region)	This is weather station data. There are many of these across the UK.
For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)	We have assumed that each set of data from the weather stations are representative of a point from which we use interpolation techniques to assign historical weather to individual fields.
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	We end up with an interpolated mean for each field for the weather variables of interest.
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	There are both hourly and daily values available from the weather stations. We use averages like mean rainfall over October to help classify the fields.
For temporal observations what is the sampling frequency.	Hourly
For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	Point values
Is uncertainty known and is this broken down into observation errors, processing errors, representativity? How is it described?	No
Do you have confidence in	We may be able to specify uncertainties stemming from any

being able to specify uncertainties? What issues can you foresee in specifying uncertainties?	of the interpolation/aggregation techniques we use in turning this data into field level information.
---	---

3.2.3 The Geophysical Data (observation component)

Data set name (unique)	Geophysical Data
Data source (where does the data come from)	This data is extracted from the LandIS - Land Information System – held and maintained by The National Soil Resources Institute (NSRI) at Cranfield University. The database covers the areas of England and Wales. http://www.landis.org.uk/index.cfm
Data availability (where can this be obtained from, access rights)	Access to this data is limited. Arrangements for access to the data are governed by an agreement between NSRI and Defra acting on behalf of the Crown. Through Defra, FERA has full access to this data. “Soilscapes” is a graphical viewer by which a summarised form of this data is displayed by NSRI, which is available to the public: http://www.landis.org.uk/services/soilscapes.cfm
Data accessibility (is this available over the web, and how is this made available?)	The data is not available over the web, and it must be protected within the web service. FERA holds this data, already extracted from the large database and matched to the fields of the JCA region, in a text-file format. FERA has access to this data for any region in England and Wales.
Data type (continuous, categorical, binary etc)	This data contains categorical and continuous variables. These are: <ul style="list-style-type: none"> • soil types (categorical) • soil texture (categorical) • drainage (categorical) • fertility (categorical) • pH value (continuous).
For spatial observations what is the domain (extent of the input region)	FERA holds this data, already matched to the field parcels of the IACS historical data for the East Anglian Chalk JCA region. The data can be obtained for other regions in England and Wales if required.
For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)	We have an observation on each variable, for each field parcel (which is defined as a polygon - see the “Geometry of Fields” observation table for further details). This is the case for a large set of field parcels of differing sizes that make up the region of interest.
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location)	For each field parcel, we have the soil type, soil texture, drainage, fertility classification and soil pH. Where more than one soil type occurs within a field parcel, the soil type that covers the largest area of the field parcel has been used. This data cleaning has already been completed for the East

value)	Anglian Chalk JCA region.
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	N/A
For temporal observations what is the sampling frequency.	N/A
For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	N/A
Is uncertainty known and is this broken down into observation errors, processing errors, representativity? How is it described?	The land-use researchers have some idea of land capability misclassification especially when considering soil type. They use some formula to go from a smoothed soil classification system to allocating the soil type for each field parcel. The data has been checked for misclassification, and therefore it will not be considered as uncertain in the model chain.
Do you have confidence in being able to specify uncertainties? What issues can you foresee in specifying uncertainties?	It is possible to formally model uncertainty in the translation from the raw soil data to the field parcel level. However, the actual classification was produced by another institution, so it will be difficult to quantify this uncertainty.

3.2.4 The Historical Crop Rotation data (observation component)

We have two sets of data here – the IACS (the Integrated Administration and Control System) agricultural land cover data and the June Survey data.

Data set name (unique)	Historical Crop Rotation Data: IACS (the Integrated Administration and Control System) agricultural land cover data. The final year of this data set (2004) will be used as our “Current Field Usage” data, which is a direct input to the field use simulator model, LandSFACTS (LandSFACTS:Inicrop).
Data source (where does the data come from)	DEFRA – The rural payments agency.
Data availability (where can this be obtained from, access rights)	Access to this data is restricted. It is extracted from the payments system which farmers use to claim subsidies. FERA hold a cleaned version of this data for the East Anglian Chalk JCA region (and some of the surrounding area of this region), in the form of a large text file, for the period 1993 – 2004.
Data accessibility (is this available over the web, and	The data is not accessible over the web. It is available internal to FERA, and must be protected within the web

how is this made available?)	service.
Data type (continuous, categorical, binary etc)	<p>The data contains continuous, discrete and categorical variables. These include:</p> <ul style="list-style-type: none"> • Spatial position information (in terms of “Easting” and “Northing”) for each field parcel. (continuous) • A reference number corresponding to each field parcel. (categorical) • Observation year. (discrete) • The area of each field parcel (in hectares, to 2dp). (continuous) <p>The land usage for each field parcel in each year – this is aggregated to 15 crop groupings. (categorical)</p>
For spatial observations what is the domain (extent of the input region)	We hold this data in a cleaned and aggregated form for the East Anglian Chalk JCA region (and some of the surrounding area of this region). The data can be obtained for other areas of England if required.
For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)	The data is categorical (a land use type) over a large set of field parcels (polygons) of different sizes that make up the region of interest.
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	<p>For each field parcel, the land use type is given – where more than one land-usage occurs within a single field parcel, the land cover type that takes the greatest area of the field is used for that field parcel. The data cleaning has been performed here as follows:</p> <p>If over 50% of the field area is used for a particular crop, then this crop is allocated to this field parcel number, and only the declared area for this crop is used. All other records of land use for the field parcel are then removed from the data set.</p> <p>If all crops within the field parcel take less than 50% of the area, then we are too uncertain about the land usage in this field parcel, and all records for that field parcel have been removed from the data set.</p>
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	We have the observed land use for each field parcel on a yearly basis. However, not all fields appear within the data in all years due to external factors, such as the system through which it was collected (farmers may not have claimed for subsidies each year), or the possibility that fields may have been sold for other land uses, eg. housing.
For temporal observations what is the sampling frequency.	Yearly
For temporal observations what is time support (the interval that one value in that	The observation of land use is categorical and hence constant for the year.

input represents, e.g. is it a time average, or instantaneous value)	
Is uncertainty known and is this broken down into observation errors, processing errors, representativity? How is it described?	<p>No. The uncertainty in this data is not known.</p> <p>The data has been cleaned. For example, if field parcels have the same spatial grid reference, they should have identical CPH number (either Main or Subsidy) no matter if the field was split or not. If there is more than one type of crop grown in the field, the data is cleaned as described above, by either using the land cover type that covers over 50% of the area of the field parcel, or by removing the records for the field parcel where the most prominent land use cannot be obtained.</p> <p>Hence, this data will not be considered as uncertain in the model.</p>
Do you have confidence in being able to specify uncertainties? What issues can you foresee in specifying uncertainties?	<p>It is possible that we could try to model the uncertainty in the data that is caused by the splitting of field parcels, and also by any mis-recording in the IACS system itself, but this system is an external system to FERA, and therefore quantifying the uncertainty would be very difficult. We believe that the method of cleaning the data described above should remove most, if not all, errors within the data set.</p>

Data set name (unique)	Historical Crop Rotation Data: June Survey Data
Data source (where does the data come from)	<p>The June Survey is an annual survey of agricultural and horticultural activity in England, run on 1st June each year. It was a full census until 1995, when it was downscaled to a sample survey. (a full census is completed every 10 years and the next is 2010)</p> <p>The survey is run by Defra, to collect detailed information on arable and horticultural cropping activities, land usage, livestock populations and agricultural labour force figures. The survey questionnaire sent out to farmers is at: http://www.defra.gov.uk/corporate/docs/forms/census/css947.pdf</p> <p>For more details, see http://www.defra.gov.uk/evidence/statistics/foodfarm/landuselivestock/junesurvey/index.htm</p> <p>This data is collected on the Farm level, rather than on the Field level.</p>
Data availability (where can this be obtained from, access rights)	<p>The full set of data at farm level is restricted, and so must be protected within the web service. The June survey data is publicly available in a summary format, which can be obtained from: http://www.defra.gov.uk/evidence/statistics/foodfarm/landuselivestock/junesurvey/index.htm</p>
Data accessibility (is this available over the web, and how is this made available?)	<p>The full set of data at farm level is not accessible over the web. It is available internal to FERA, and so must be protected within the web service. The June survey data is</p>

	<p>publicly available in a summary format, which can be obtained from:</p> <p>http://www.defra.gov.uk/evidence/statistics/foodfarm/landuselivestock/junesurvey/index.htm</p> <p>The format of these summaries is tabular within a .pdf report document, and regional and county level statistics can be gained in the form of excel spreadsheet.</p>
<p>Data type (continuous, categorical, binary etc)</p>	<p>The data we have from the June survey is given on the farm level. It contains the total area of farmed land for the holding, and then also this area broken down into the different totals of the area covered by each land use type on the farm. (area is in hectares)</p>
<p>For spatial observations what is the domain (extent of the input region)</p>	<p>The survey considers land use at the farm level, and covers the area of England.</p>
<p>For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)</p>	<p>To enable June Survey data to be grouped into different geographic areas, such as region or county, every holding is allocated a grid reference. A number of steps are followed to do this.</p> <p>1. Firstly the average grid reference of all of the fields on the holding is calculated. Then, in order of preference, the average grid reference is tested against the following conditions in order to select which field should be used to set the holding’s grid reference:</p> <ul style="list-style-type: none"> • If the grid reference of the field closest to the average is in the same parish as the holding then this is chosen. • If the field nearest the average grid reference point is not in the same parish as the holding but it is in the same ward this is used. • If the average field grid reference is not in the same ward or parish as the holding but is within 10 kilometres of the central point of the holding’s parish then the field grid reference closest to this central point will be used. <p>If the average field grid reference is not in the same ward or parish as the holding but is within 15 kilometres of the central point of the holding’s parish the field grid reference closest to this central point will be used.</p> <p>2. If still no grid reference can be found the contact address for the holding is looked at. It is possible to generate a grid reference from the postcode. First the postcode is checked to make sure it is in the same parish or ward to which the farm has been allocated, as the contact address is not always at the same location as the farm.</p> <p>3. For the remaining grid references checks are run to see if the holdings have ever been given a grid reference in previous years and if it has this is then allocated.</p>

	4. Finally, for those holdings where no grid reference can be found a random grid reference is allocated to them from the same parish to which the farm has been allocated. In 2008 approximately 9 per cent of holdings had grid references generated this way.
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Each line of observations represents a farm holding.
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	This is a yearly survey.
For temporal observations what is the sampling frequency.	This is a yearly survey. Although, this is done using a stratified sampling method and so some of the sampled holdings may differ between years. The strata are set up to ensure a full mix of large and small holdings and different cropping types are included into the sample. In 2010, all holdings will be sampled and the survey will take the form of a census.
For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	The observations each year are single observations that are set for the year period (areas of land use)
Is uncertainty known and is this broken down into observation errors, processing errors, representativity? How is it described?	<p>Uncertainty is considered on the outputs from the survey. These uncertainties include:</p> <ul style="list-style-type: none"> • The sampling rate of the survey is around 20% of the agricultural population – estimates have to be made for non-sampled/ non-responding holdings (using an imputation procedure?). Indicators are published alongside estimates derived from the survey data to give an indication of the scale of the sampling errors. • It is a postal survey, so there is a degree of non-response which may potentially cause a bias in the results. The response rate is still reasonably high at 70% and how the response differs between farm types and size is monitored to try and avoid this possibility. • As this is a postal survey, the data is also subject to the vagaries of farmer's interpretation of the categories. Defra try to keep the form as clear as possible, to try and minimise confusion. A data validation exercise to clean the data prior to processing also helps to keep the data as accurate as possible.

Do you have confidence in being able to specify uncertainties? What issues can you foresee in specifying uncertainties?	No.
---	-----

3.2.5 The Land Capability Classification System (Model Component).

General information about the model component:

Model name (make this unique)	The Land Capability Classification System (LCCS)
Model creator (where does it come from?)	FERA
Model licence (who can use it? Is it open source?)	FERA
Model requirements (operating system, hardware)	Runs on a desktop PC under Windows operating system. Uses the statistical software R, which can be run on most operating systems.
Model computation (typical time to evaluate given a set of typical inputs – may require some commentary)	Computation time is dependent on the number of land capability classifications used and the size of the historical crop rotation data set. For the East Anglian Chalk JCA region, computation time is minimal.
Model description (<i>overview</i> of what the model does, how it works – this will need to be brief)	<p>This model takes climate data (either current or some simulated future climate scenario gained from the climate model), geophysical data (such as soil type and field drainage) and the IACS historical crop rotation data for the fields in the region under study to group the field parcels by land capability and hence determine which types of crops can be grown in each field.</p> <p>The model outputs the historical crop rotation data in sets corresponding to the different land capabilities of interest, which are chosen by the user.</p>
Relevant references (papers or other resources, including web links, that describe the model)	<p>FERA internal report for a recent land use project:</p> <p>W. Luo and N. Boatman (2010): Modelling agricultural land use, with a focus on crop rotation.</p>
Website or link for downloading the model, if available.	This is an internal model to FERA and therefore no such links are available.
Additional comments	None.

Summary of component model inputs:

Input name	Short description	Uncertain?	Importance
------------	-------------------	------------	------------

			rank
IACS historical crop rotation data	Land use data for the field parcels in the region of interest over the period 1993 – 2004 (described in the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table.)	Possibly – See observation table.	1
Geophysical data	Data on soil types, texture, drainage, fertility and pH value for the field parcels in the area under study (see the “Geophysical Data” observation table for more details)	Possibly – See observation table.	1
Climate data	Data on climate variables such as rainfall and temperature at different site (weather station) locations over the region of interest (see the “Climate Data” observation table for more details).	Possibly – See observation table.	5
Land capability classification rules (rules)	The rules by which we classify a field parcel as having a certain land capability.	Yes	1

Description of uncertain model inputs:

Input name (unique if possible – add model name e.g. model:input)	LCCS:rules
Input role (parameter in model, initial condition, boundary condition)	This input provides the rules by which we classify land capability over the region of interest.
Input type (continuous, categorical, binary etc)	Categorical.
For spatial inputs what is the domain (extent of the input region)	N/A
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	N/A
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	N/A
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	N/A
For temporal inputs what is	N/A

the time resolution (and how is this represented – time series, spectral, other)	
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	N/A
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	No, uncertainty is not known. This input is determined by the user. The user may not be sure on the classification rules to use, especially in terms of the boundaries of classifications, for example: The boundaries to use when breaking down the soil pH variable into different classes. This uncertainty may be quite difficult to quantify.
Is the marginal (at a single location or object) probability density function known for this input?	N/A
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	N/A
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	N/A
Do you know, and can you specify typical ranges for this input?	For the East Anglian Chalk JCA region, if the “soilscapes” soil type only is used for the classification, then we will have 12 different land capabilities. These classifications can be broken down into different soil textures, drainage and fertility, which could enable more or less groupings to be used. Variables such as climate behaviour could also be used to determine a set of land capability classes, which can allow the user to classify the land capability in different ways.
Input source (describe where this input might come from – link to a observation component if this makes sense)	This input is determined by the user – a default can be taken as the “soilscapes” classifications for soil types.

Summary of component model outputs:

Output name	Short description	Importance rank
LCdatasets	Sets of crop rotation data corresponding to the different land capabilities under	1

	consideration.	
Land capability classifications (LCclass)	A list of the land capability classifications that correspond to the sets of data in LCdatasets.	1
The land capability allocation by field (LCfieldalloc)	A table containing: <ul style="list-style-type: none"> • The IACS field reference number, and • The allocated land capability classification corresponding to that field, for all field parcels in the region of interest.	1

Descriptions of uncertain model outputs:

Output name in form model:output	LCCS:LCdatasets
Output role (feeds into / is used for)	This output feeds into the Crop Allocation Mechanism (CAM)
Output type (continuous, categorical, binary etc)	This data output contains continuous, discrete and categorical variables – the same variables as those outlined in the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table. Here, the data for each field parcel has been allocated to a set depending on the land capability of the field, where each field has been classified using the geophysical data and potentially also the climate data. The set of classifications are based on these data sets, and are contained in the list: LCclass
For spatial outputs what is the domain (extent of the input region)	Each set of field parcel level crop data within this output contains data for field parcels within the region of interest – initially we will concentrate on the East Anglian Chalk JCA region (and some of the surrounding area of this region).
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	Each data set contains a land use type (categorical) for each field parcel within the region of interest (represented as a polygon – see the “Geometry of Fields Data” observation table) that has the corresponding land capability.
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	This is the same as that described for the IACS historical crop rotation data, in the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table.
For temporal outputs what is the domain (extent of the	The data is yearly, although not all fields occur within the data for each year (see the “Historical Crop Rotation Data:

input time – at a point, over a time interval?)	IACS agricultural land cover data” observation table).
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	Yearly
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	The observation of land use is categorical and hence constant for the year.
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble, distribution)?	No, the uncertainty in this data is not known – see the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table. Also, uncertainty may be induced in the allocation to land capability, as some field parcels may have more than one soil type/land capability. As explained in the “Geophysical Data” observation table, the dominant soil type that covers the largest area within the field parcel has been used as the soil type for the full field parcel, and there may be uncertainty associated within this procedure.
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	No
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support.	N/A – this output contains the data observations.
Will visualisation be necessary? How might this be done?	No
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	No – this output contains the data observations.
Output name in form model:output	LCSS:LCclass
Output role (feeds into / is used for)	This output feeds directly alongside the output LCSS:LCdatasets into the Crop Allocation Mechanism (CAM), and also alongside the output CAM: CTM from the crop Allocation Mechanism, into the field use simulator model, LandSFACTS. It is the labelling for the sets of data

	and the transition matrices, to show which one corresponds to which land capability.
Output type (continuous, categorical, binary etc)	Categorical
For spatial outputs what is the domain (extent of the input region)	N/A
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	N/A
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	N/A
For temporal outputs what is the domain (extent of the input time – at a point, over a time interval?)	N/A
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	N/A
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	N/A
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble, distribution)?	N/A. This is a list of the land capability classifications that correspond to the sets of data in LCdatasets. These classifications are fixed.
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	The rules for the classification are entered by the user as the model input LCCS:rules, and so model error in the different classifications of land capability may exist through user interpretation of the boundaries (see the input table for LCCS:rules). Hence it is possible that, based on the same information, different users could generate different sets of classifications.
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are	N/A

there issues of differing support.	
Will visualisation be necessary? How might this be done?	No
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	N/A

Output name in form model:output	LCCS:LCfieldalloc
Output role (feeds into / is used for)	This output feeds directly into the field use simulator model, LandSFACTS, and corresponds to the input: LandSFACTS:Fclass
Output type (continuous, categorical, binary etc)	Categorical.
For spatial outputs what is the domain (extent of the input region)	This output contains the land capability classification for each field parcel within the region of interest – initially the East Anglian Chalk JCA region (and some of the surrounding area of this region).
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	The region is represented as field parcels of differing size (see the “Geometry of Fields Data” observation table).
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	This output contains the corresponding land capabilities for all the field parcels in the region of interest.
For temporal outputs what is the domain (extent of the input time – at a point, over a time interval?)	N/A
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	N/A
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	N/A
Is output uncertainty currently known and	No, the uncertainty is not currently known. It is possible that uncertainty may be induced in the allocation to land

quantified and to what degree? How is it described (samples, ensemble, distribution)?	capability, as explained in the output table for LCCS:LCdatasets.
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	No, this is not known for this output. It is possible that misclassification could occur in allocating the land capability to the field parcels, as described above in the output table for LCCS:LCclass.
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support.	No
Will visualisation be necessary? How might this be done?	Yes. We would like to be able to map the classifications over the region. We would also need the information about the classifications in LCCS:LCclass and the Geometry of Fields Data to do this.
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	No.

3.2.6 The Markovian Crop Allocation Mechanism (model component).

General information about the model component:

Model name (make this unique)	The Crop Allocation Mechanism (CAM)
Model creator (where does it come from?)	FERA.
Model licence (who can use it? Is it open source?)	FERA.
Model requirements (operating system, hardware)	Runs on a desktop PC under Windows operating system. Uses the statistical software R, which can be run on most operating systems.
Model computation (typical time to evaluate given a set of typical inputs – may require some commentary)	Computation time is minimal (less than 10 secs per transition matrix for the East Anglian Chalk JCA region), although this does depend on the number of land capability classes (and hence the number of transition matrices to be produced), and the amount of historical crop rotation data being used to compute the matrices.
Model description (<i>overview</i> of what the model does, how it works – this will need to be brief)	This model uses the historical crop rotation data grouped by land capability, and possibly economic constraints, to determine a matrix of transition probabilities for crop rotation under each land capability, for the field parcels in an

	area under study. It is a statistical methodology, based on transition matrices in a Markov chain. Different climate and economic scenarios can affect the model.
Relevant references (papers or other resources, including web links, that describe the model)	FERA internal report for a recent land use project: W. Luo and N. Boatman (2010): Modelling agricultural land use, with a focus on crop rotation.
Website or link for downloading the model, if available.	This is an internal model to FERA and therefore no such links are available.
Additional comments	None.

Summary of component model inputs:

Input name	Short description	Uncertain?	Importance rank
LCdatasets	Grouped sets of data from the IACS historical crop rotation data, generated through the Land Capability Classification System (LCCS). Each set of data within LCdatasets corresponds to the IACS historical crop rotation data of fields that are classed as having a particular land capability.	Yes	1
LCclass	This is the list of the land capability classifications that correspond to the sets of data in LCdatasets.	No	1

Description of uncertain model inputs:

Input name (unique if possible – add model name e.g. model:input)	CAM:LCdatasets
Input role (parameter in model, initial condition, boundary condition)	This is the data from which we generate a crop rotation transition matrix for each land capability in the classification list generated by the Land Capability Classification System, outputted in LCCS: LCclass.
Input type (continuous, categorical, binary etc)	<p>This is a data input, which contains continuous, discrete and categorical variables – the same variables as those outlined in the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table.</p> <p>The main variables used here are:</p> <ul style="list-style-type: none"> • the reference number corresponding to each field parcel (categorical) • the observation year (discrete) • the land usage for each field parcel in each year – this is aggregated to the 15 crop groupings (categorical) <p>See also, LCCS: LCdatasets</p>

For spatial inputs what is the domain (extent of the input region)	See output: LCCS:LCdatasets.
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	See output: LCCS:LCdatasets.
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	See output: LCCS:LCdatasets.
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	See output: LCCS:LCdatasets.
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	See output: LCCS:LCdatasets.
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	See output: LCCS:LCdatasets.
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	No, this is not known – See output: LCCS:LCdatasets.
Is the marginal (at a single location or object) probability density function known for this input?	N/A
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	N/A
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	N/A

Do you know, and can you specify typical ranges for this input?	The year variable takes values between 1993 and 2004. The land use for each field parcel is aggregated to a high level classification that has 15 crop groupings. Not all field parcels are available in all years (see the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table)
Input source (describe where this input might come from – link to a observation component if this makes sense)	This input comes from the output of the Land Capability Classification System: LCCS: LCdatasets. It also links to the IACS historical crop rotation data described in the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table.

Summary of component model outputs:

Output name	Short description	Importance rank
Crop Transition Matrices (CTM)	A list containing crop transition matrices for each of the land capabilities. For each set of data in LCdatasets, and therefore each land capability in LCCS:LCclass, CTM holds a corresponding matrix of transition probabilities for crop rotation – for moving from the crop type of the row to the crop type of the column.	1

Description of uncertain model output:

Output name in form model:output	CAM:CTM
Output role (feeds into / is used for)	This output feeds into the field use simulator model, LandSFACTS, as the input: LandSFACTS:Tmats.
Output type (continuous, categorical, binary etc)	The output is continuous on the scale [0,1]: a set of matrices of probabilities.
For spatial outputs what is the domain (extent of the input region)	N/A
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	N/A
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	N/A
For temporal outputs what is the domain (extent of the	N/A – Each matrix is considered to be the transition matrix of a stationary distribution, and hence not changing through

input time – at a point, over a time interval?)	time.
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	N/A
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	N/A
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble, distribution)?	No, this uncertainty is not currently known and quantified. The crop transition matrices are constructed empirically from the historical data, and therefore uncertainty is present in the transition probabilities. For land capabilities of which the size of the field sample is low, then this uncertainty is higher. The uncertainty in this output could be modelled using Dirichlet distributions over the rows of the transition matrices.
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	No.
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support.	No.
Will visualisation be necessary? How might this be done?	No – but if required, simple graphical output such as bar charts would suffice.
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	This output is observation driven.

3.2.7 The Economic Data (observation component).

Data set name (unique)	Economic Data
Data source (where does the data come from)	This data is taken from the “Farm Management Pocketbook” for each year, up to 2009.
Data availability (where can this be obtained from, access)	The “Farm Management Pocketbook” is a yearly publication (written by John Nix) which details forecasts for yields and crop prices for the next coming year, along with the pricing

rights)	of other variable agricultural costs such as fertilizers and labour.
Data accessibility (is this available over the web, and how is this made available?)	This is a yearly publication (book) and is publically available. FERA holds this publication for many years up to 2009.
Data type (continuous, categorical, binary etc)	Continuous (Projected crop prices - £, per hectare.)
For spatial observations what is the domain (extent of the input region)	N/A
For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)	N/A
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	N/A
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	N/A
For temporal observations what is the sampling frequency.	N/A
For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	N/A
Is uncertainty known and is this broken down into observation errors, processing errors, representativity? How is it described?	No. The values given are average prices. The values are given for “Low”, “Average” and “High” performance and production levels for a farm. These are meant to indicate differences in natural factors, soil productivity, and/or managerial skill of the farmer. Using information collected in the Defra June Survey (see the “Historical Crop Rotation Data: June Survey Data” observation table for more details), we may also be able to link this economic data with Farm types. We are not entirely sure how/if we can incorporate this information into the model chain – this is still under review.
Do you have confidence in being able to specify	No

uncertainties? What issues can you foresee in specifying uncertainties?	
---	--

3.2.8 The Geometry of the Fields (observation component)

Data set name (unique)	Geometry of Fields Data
Data source (where does the data come from)	This data is held by FERA. It is taken from the Rural Land Register (RLR) spatial database, which holds digital mapping information for all the land receiving subsidies.
Data availability (where can this be obtained from, access rights)	<p>FERA holds this data. Using GIS applications, the spatial make-up of the field parcels from RLR has been mapped to the field parcels within the IACS historic crop rotation data. Hence, the geography of the fields for the East Anglian Chalk JCA region is in the form of a GIS shape file that contains polygons defining the field parcels. Each field parcel has the corresponding IACS field code attached.</p> <p>This data is sensitive (fields could be identified as belonging to a particular holding) and therefore access to the data behind the shape polygons (IACS codes etc) is restricted, and must be protected within the web service.</p>
Data accessibility (is this available over the web, and how is this made available?)	This data is not available over the web. FERA has full access to the data. Being restricted, this data must be protected within the web service.
Data type (continuous, categorical, binary etc)	The data is in the form of a GIS shape file, which can be loaded directly into the land-use simulator model, LandSFACTS.
For spatial observations what is the domain (extent of the input region)	This data is already in the GIS format ready for use for the East Anglian Chalk JCA region. We can obtain this data for other areas of England if required.
For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)	The field parcels are represented as polygons within this data source. The field parcels all differ in size and shape.
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Each polygon within the shape file corresponds to a field parcel within our region, for which we have a current land use type and historical data.
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	N/A
For temporal observations what is the sampling frequency.	N/A

For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	N/A
Is uncertainty known and is this broken down into observation errors, processing errors, representativity? How is it described?	The geography of the fields for this region has been previously checked against “Google” maps of the areas to pick up any irregularities. We therefore consider this data as fixed and do not consider uncertainty here.
Do you have confidence in being able to specify uncertainties? What issues can you foresee in specifying uncertainties?	N/A

3.2.9 The Field Use Simulator – LandSFACTS Model

General information about the model component:

Model name (make this unique)	LANDscape Scale Functional Allocation of Crops Temporally and Spatially (LandSFACTS)
Model creator (where does it come from?)	The Macaulay Land Use Research Institute, Aberdeen, UK.
Model licence (who can use it? Is it open source?)	The software is released under GNU public license, and is therefore free to use and develop. The code sources are available upon request.
Model requirements (operating system, hardware)	It runs on a desktop PC under the Windows operating system (however, I think that it will work on other systems with some tweaks to the source code).
Model computation (typical time to evaluate given a set of typical inputs – may require some commentary)	This is highly dependent on the number of simulations being run. We run a number of simulations of the model as a big stochastic element. The time needed also depends on the area being simulated: the more fields there are, the greater the number of stochastic choices. For a similar application to the ones we will be running in UncertWeb, the program took about 5 hours to run for 1,000 simulations for a given set of inputs (on a desktop PC).
Model description (<i>overview</i> of what the model does, how it works – this will need to be brief)	The model takes information about possible crop rotations, land capability, current field usage and temporal/spatial constraints on cropping and simulates cropping patterns across each field for a given number of years. This is achieved through stochastic simulation of field usage using different transition matrices for different land capabilities (all within the specified constraints).
Relevant references (papers or other resources, including	http://www.macaulay.ac.uk/LandSFACTS/ Castellazzi, M. S., Matthews, J., Wood, G. A., Burgess, P. J.,

web links, that describe the model)	<p>Conrad, K. F., and Perry, J. N. (2007). LandSFACTS: Software for Spatio-temporal Allocation of Crops to Fields Proceedings of 5th Annual Conference of the European Federation of IT in Agriculture.</p> <p>Castellazzi, M. S. (2007). Spatio-temporal modelling of crop-coexistence in European agricultural landscapes. PhD Thesis, Cranfield University.</p>
Website or link for downloading the model, if available.	http://www.macauley.ac.uk/LandSFACTS/download.php
Additional comments	We have secured the source code from the developers. The internal stochastic calculations are implemented in a Python code. The model front-end is written in C++. In its current form, the model requires a lot of front-end interaction in order to start a simulation. We are exploring ways of circumnavigating this.

Summary of component model inputs:

Input name	Short description	Uncertain?	Importance rank
Field polygons (Fshape)	A GIS shape file describes the locations of all the fields under consideration. This input corresponds to the Geometry of Fields data.	No	1
Field classification (Fclass)	Each of the fields in the model must have an associated land capability class so that the appropriate rotation transition matrix can be used. This corresponds to the output LCCS:LCfieldalloc from the Land Capability Classification System (LCCS).	Yes	1
Crops (Crops)	The model needs to know what crop types we are considering in the simulation. We will probably use 15 broad classes of crops, which line up to Defra's definitions.	No	3
Transition matrices (Tmats)	For each Fclass, there is a transition matrix of probabilities for the updating of the fields from year-to-year. This corresponds to the output CAM:CTM from the Crop Allocation Mechanism (CAM).	Yes	1
Initial crop allocation (Inicrop)	Each field needs a starting point for the simulation: these will be based on actual field level data. We intend to use the final year (2004) from the IACS Historical Crop Rotation Data for this input (see the "Historical Crop Rotation Data: IACS agricultural land cover data" observation table for more details on this data).	Yes	4

Tolerable level of cropping (Tprop)	For some types of crops, we might want to add a limit to how many fields are allocated. For instance, there could be a limit on the proportion of GM crops grown.	No	4
Spatial constraints (Scons)	Some crops must be kept separate.	No	8
Temporal constraints (Tcons)	Some crops cannot be grown in consecutive years.	No	4
Length of simulation (Nyears)	This dictates how many years into the future the simulation runs.	No	5
Tolerance levels (Tol)	As this is a stochastic simulation under constraints, the model searches for allowable outcomes. The tolerance levels dictate the flexibility of these searches.	No	8
Number of repeated simulations (Nsim)	This parameter controls the number of stochastic realisations the model produces.	No	3

Description of uncertain model inputs:

Input name (unique if possible – add model name e.g. model:input)	LandSFACTS:Fclass. This input corresponds to the output LCCS:LCfieldalloc from the Land Capability Classification System (LCCS).
Input role (parameter in model, initial condition, boundary condition)	This input describes the land capability of each field parcel within the region of interest.
Input type (continuous, categorical, binary etc)	Categorical
For spatial inputs what is the domain (extent of the input region)	See the output: LCCS:LCfieldalloc.
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	See the output: LCCS:LCfieldalloc.
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	See the output: LCCS:LCfieldalloc.

For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	N/A
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	N/A
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	N/A
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	The land-use researchers have some idea of land capability misclassification especially when considering soil type. They use some formula to go from a smoothed soil classification system to allocating soil type for each field. The measures of uncertainty are considered with the output LCCS:LCfieldalloc.
Is the marginal (at a single location or object) probability density function known for this input?	No
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	No
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	No
Do you know, and can you specify typical ranges for this input?	See the outputs LCCS:LCfieldalloc and LCCS:LCclass.
Input source (describe where this input might come from – link to a observation component if this makes sense)	This input corresponds to the output from the Land Capability Classification System: LCCS:LCfieldalloc.

Input name (unique if possible – add model name e.g. model:input)	LandSFACTS:Tmats. This input corresponds to the output CAM:CTM from the Crop Allocation Mechanism (CAM).
Input role (parameter in	This input refers to a set of transition matrices that govern

model, initial condition, boundary condition)	the stochastic part of the crop allocation within the model. There is a set of matrices as a matrix is needed for each class of field.
Input type (continuous, categorical, binary etc)	Continuous. This input is a set of matrices whose elements are probabilities and rows sum to one.
For spatial inputs what is the domain (extent of the input region)	N/A
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	N/A
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	N/A
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	N/A – See the output CAM:CTM
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	N/A
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	N/A
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	See the output CAM:CTM.
Is the marginal (at a single location or object) probability density function known for this input?	See the output CAM:CTM.
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	See the output CAM:CTM.
Is the joint (across multiple	See the output CAM:CTM.

inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	
Do you know, and can you specify typical ranges for this input?	The elements of the matrices are probabilities. Experts in land use should have a fair idea of what the chances are of changing between the major crop types.
Input source (describe where this input might come from – link to a observation component if this makes sense)	The transition matrices correspond to the output from the Crop Allocation Mechanism (CAM).

Input name (unique if possible – add model name e.g. model:input)	LandSFACTS:Inicrop. This input comes from the IACS Historical Crop Rotation Data.
Input role (parameter in model, initial condition, boundary condition)	This is a list of crops that are present in the field at the beginning of the simulation period.
Input type (continuous, categorical, binary etc)	Categorical (the number of options depend on LandSFACTS:Crops) for each field.
For spatial inputs what is the domain (extent of the input region)	See the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table.
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	See the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table.
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	See the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table.
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	This input represents the land use for the field parcels in the region of interest for one year (the initial year)
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	Yearly.
For temporal inputs what is time support (the period that	One year.

one value in that input represents, e.g. time average or instantaneous value)	
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	If the simulations are started from present (or close enough to present), then we have a very good idea of what crops are in which farms. Of course, we may need to make assumptions about how they are distributed across the fields of the farm if we have farm level constraints. Further details are in the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table.
Is the marginal (at a single location or object) probability density function known for this input? Where will this come from (expert elicitation, using data, instrumental error, interpolation error, classification error)?	No. Currently, the best estimate we have of current field level usage is taken as being known. However, we might be able to say something about classification errors. There is an option within the program to randomly generate the crop allocations prior to the simulation. For simulations starting in the future, we have less of a handle on what might be appropriate for this initial condition.
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	Same as above.
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	Same as above.
Do you know, and can you specify typical ranges for this input?	We can talk about typical proportions of crops for different land capabilities under current climatic and economic conditions.
Input source (describe where this input might come from – link to a observation component if this makes sense)	We intend (initially) to use the final year (2004) from the IACS Historical Crop Rotation Data for this input (see the “Historical Crop Rotation Data: IACS agricultural land cover data” observation table for more details on this data).

Summary of component model outputs:

Output name	Short description	Importance rank
Area of fields (FArea)	For a set of field polygons, the model calculates the area of each field so that it can be used to calculate areas used for each crop type. (This is not variable because we are not considering uncertainty in field geometry.)	10
Area of	For each class of land capability, the area of	1

cropping (CropArea)	land used for each crop is given along with the proportion of land for each crop.	
Simulated rotations (SimRot)	For each field and for every simulated year, the model allocates a cropping choice.	5

Description of uncertain model output:

Output name in form model:output	LandSFACTS:FArea.
Output role (feeds into / is used for)	This output is a by-product of LandSFACTS:CropArea. The model calculates the area of each field so we can say how much land falls under each land capability.
Output type (continuous, categorical, binary etc)	Continuous (areas in hectares).
For spatial outputs what is the domain (extent of the input region)	N/A
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	N/A
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	N/A
For temporal outputs what is the domain (extent of the input time – at a point, over a time interval?)	N/A
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	N/A
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	N/A
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble,	This output is taken to be a precise measurement of field area.

distribution)?	
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	N/A
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support.	N/A
Will visualisation be necessary? How might this be done?	No
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	Some of the field polygons have been checked against aerial photographs of the corresponding farmland.

Output name in form model:output	LandSFACTS:CropArea.
Output role (feeds into / is used for)	This output feeds into the yield model to help predict the amount of each crop type that is produced over the region of interest.
Output type (continuous, categorical, binary etc)	This output is continuous (areas in hectares), but bounded by the total FArea and zero for each crop type.
For spatial outputs what is the domain (extent of the input region)	The region of interest. Initially, we will concentrate on the East Anglian Chalk JCA region.
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	This output is aggregated to the level of the full region of interest – we have a single value for the region, for the area of land covered by each crop.
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	The full region.
For temporal outputs what is the domain (extent of the input time – at a point, over a time interval?)	The crop area is produced for each year of the simulation.
For temporal outputs what is	Time series stepping over simulated years.

the time resolution (and how is this represented – time series, spectral, other)	
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	A year.
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble, distribution)?	The uncertainty in this output stems from the stochastic nature of the model. Currently, this is shown through a sample of outputs from the model.
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	No.
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support. Please provide a complete description of the observations as a data component	This depends on the starting point we use for the simulation. If we start with the most recent available data, we have no observations available. However, we could perform history matching by starting the model in 2000 say. Regional estimates of crop area could be obtained from the June Survey data (see the “Historical Crop Rotation Data: June Survey data” observation table), which could prove useful for model validation. (Especially at the end of the current year – 2010 – as the survey this year is a full census, and so should provide an accurate estimate for the current year when this data becomes available.)
Will visualisation be necessary? How might this be done?	Not of this aggregated output.
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	No.

Output name in form model:output	LandSFACTS:SimRot.
Output role (feeds into / is used for)	This output is used to visualise evolutions in cropping over the simulated years.
Output type (continuous, categorical, binary etc)	Categorical.
For spatial outputs what is the domain (extent of the input)	There is a simulated crop for each field parcel in the region of interest.

region)	
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	Field level.
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Each crop value belongs to one specific field parcel.
For temporal outputs what is the domain (extent of the input time – at a point, over a time interval?)	There is a crop for each year of the simulation.
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	There is an assumption of only one crop per year, per field parcel.
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	One value represents one year's crop.
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble, distribution)?	This output is the result of a stochastic choice model. The uncertainty is shown through a sample of outputs from the model.
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	No.
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support.	See the output LandSFACTS:CropArea.
Will visualisation be necessary? How might this be done?	Yes. The land-use experts like to produce GIS graphics that show the region changing over time.
Is this output currently validated (in the sense of comparing to observations)?	No.

How is this currently done?	
-----------------------------	--

3.2.10 Expected Yield Data (observation component)

We have potential yield data from two sources: The yearly publication “Farm Management Pocketbook”, and Defra.

Data set name (unique)	Expected Yield Data – Farm Management Pocketbooks.
Data source (where does the data come from)	This data is taken from the “Farm Management Pocketbook” for each year, up to 2009.
Data availability (where can this be obtained from, access rights)	The “Farm Management Pocketbook” is a yearly publication (written by John Nix) which details forecasts for yields and crop prices for the next coming year, along with the pricing of other variable agricultural costs such as fertilizers and labour.
Data accessibility (is this available over the web, and how is this made available?)	This is a yearly publication (book) and is publically available. FERA holds this publication for many years up to 2009. We can easily transform this to an electronic format.
Data type (continuous, categorical, binary etc)	Continuous – in “tonnes per hectare”.
For spatial observations what is the domain (extent of the input region)	N/A
For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)	N/A
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	N/A
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	Yearly.
For temporal observations what is the sampling frequency.	Yearly.
For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	A Year.
Is uncertainty known and is	No. The values given are average yields. The values are

<p>this broken down into observation errors, processing errors, representativity? How is it described?</p>	<p>given for “Low”, “Average” and “High” performance and production levels for a farm. These are meant to indicate differences in natural factors, soil productivity, and/or managerial skill of the farmer.</p>
<p>Do you have confidence in being able to specify uncertainties? What issues can you foresee in specifying uncertainties?</p>	<p>No.</p>

<p>Data set name (unique)</p>	<p>Expected Yield Data – Defra</p>
<p>Data source (where does the data come from)</p>	<p>This data is available to FERA from Defra. It contains different average yield values for different areas of the UK. These yield values are calculated from information collected in the “June Agricultural Survey“, run by Defra and described in the “Historical Crop Rotation Data: June Survey data” observation table.</p>
<p>Data availability (where can this be obtained from, access rights)</p>	<p>This data is in the form of summary statistics and can be obtained from the Defra website for the years 2005 - 2009: http://www.defra.gov.uk/evidence/statistics/foodfarm/food/cereals/documents/cps_osr_mincrop_final.pdf This data is in the public domain. FERA holds this data for a much longer period of time, in the form of an Excel Spreadsheet, which is not in the public domain. Hence, access to this set of data is restricted and it must be protected within the web service.</p>
<p>Data accessibility (is this available over the web, and how is this made available?)</p>	<p>As above.</p>
<p>Data type (continuous, categorical, binary etc)</p>	<p>Continuous – in “tonnes per hectare”.</p>
<p>For spatial observations what is the domain (extent of the input region)</p>	<p>N/A</p>
<p>For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)</p>	<p>N/A</p>
<p>For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)</p>	<p>N/A</p>

For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	Yearly.
For temporal observations what is the sampling frequency.	Yearly.
For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	A Year.
Is uncertainty known and is this broken down into observation errors, processing errors, representativity? How is it described?	For the publicly available data, 95% confidence bounds are given for the average yield values.
Do you have confidence in being able to specify uncertainties? What issues can you foresee in specifying uncertainties?	We may be able to use these 95% confidence bounds, along with additional expert opinion, to describe uncertainties in this data.

3.2.11 The Yield Model

This model outputs the final output of the model chain: The Regional estimates of Future Crop Production.

General information about the model component:

Model name (make this unique)	Yield Model (YIELD)
Model creator (where does it come from?)	FERA
Model licence (who can use it? Is it open source?)	FERA
Model requirements (operating system, hardware)	Runs on a desktop PC under Windows operating system. Uses the statistical software R, which can be run on most operating systems.
Model computation (typical time to evaluate given a set of typical inputs – may require some commentary)	Computation time is minimal, as this model is not complex.
Model description (<i>overview</i> of what the model does, how it works – this will need to be	This model takes the simulated area for each crop from the output of the LandSFACTS field use simulator model, LandSFACTS:CropArea, and multiplies this by the average

brief)	expected yield from the expected yield data set.
Relevant references (papers or other resources, including web links, that describe the model)	FERA internal report for a recent land use project: W. Luo and N. Boatman (2010): Modelling agricultural land use, with a focus on crop rotation.
Website or link for downloading the model, if available.	This is an internal model to FERA and therefore no such links are available.
Additional comments	There are large yield models and yield simulators available, (for example, the “Sirius wheat simulation model” (http://www.rothamsted.bbsrc.ac.uk/mas-models/sirius.php), developed at Rothamsted), but these tend to be rather complex and have many inputs such as farmer behaviour and managerial skill, fertilizers used, soil productivity and climate. We do not have all the required information to populate such models, and therefore we take a much simpler approach to model yield, as described.

Summary of component model inputs:

Input name	Short description	Uncertain?	Importance rank
The areas of land covered by each crop. (CropAreas)	This input contains the total area in the region of interest that is covered by each land use type in each simulated year. It is the output LandSFACTS:CropArea from the field use simulator model, LandSFACTS.	Yes	1
Expected Yield Data (ExYDat)	This is the average expected yield data for each crop – possibly also broken down by area of the UK (as described in the “Expected Yield Data” observation tables).	Yes – See the “Expected Yield Data” observation tables.	1

Description of uncertain model inputs:

Input name (unique if possible – add model name e.g. model:input)	YIELD:CropAreas
Input role (parameter in model, initial condition, boundary condition)	This is a data input, which contains the area of land that has been allocated to each of the 15 aggregated crop (land use) types, for each year of the simulation from the LandSFACTS Model.
Input type (continuous, categorical, binary etc)	Continuous (these are areas in hectares).
For spatial inputs what is the domain (extent of the input region)	See the output: LandSFACTS:CropArea.

For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	See the output: LandSFACTS:CropArea.
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	See the output: LandSFACTS:CropArea.
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	This input contains the crop area for each year of the LandSFACTS simulation.
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	It is Time series stepping over simulated years for each land use type.
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	A year.
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	See the output: LandSFACTS:CropArea.
Is the marginal (at a single location or object) probability density function known for this input?	See the output: LandSFACTS:CropArea.
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	No
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	No
Do you know, and can you specify typical ranges for this input?	Yes. These are driven by the output from the field use simulator model, LandSFACTS, and the size of the region of interest. The areas are in hectares.
Input source (describe where	This input comes from the output: LandSFACTS:CropArea,

this input might come from – link to a observation component if this makes sense)	from the field use simulator model, LandSFACTS.
---	---

Summary of component model outputs

Output name	Short description	Importance rank
Regional Estimates of Future Crop Production. (RegCropEsts)	The output here is an estimate for the production / yield of each crop over the region of interest, for each simulated year from the field use simulator model, LandSFACTS. This is the overall output from our model chain.	1

Description of uncertain model output:

Output name in form model:output	YIELD: RegCropEsts
Output role (feeds into / is used for)	This is the final output from our model chain. It is an estimate of the production / yield of each crop over the region of interest, for each simulated year from the field use simulator model, LandSFACTS.
Output type (continuous, categorical, binary etc)	Continuous (tonnes).
For spatial outputs what is the domain (extent of the input region)	The region of interest. Initially, we will concentrate on the East Anglian Chalk JCA region.
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	This output is aggregated to the level of the full region of interest – we have a single value for the region, for the estimated total yield of each crop.
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	The full region.
For temporal outputs what is the domain (extent of the input time – at a point, over a time interval?)	The estimates are made for each simulated year gained from the field use simulator model, LandSFACTS.
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	Yearly, for the simulated years from the field use simulator model, LandSFACTS.

For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	The crop production for each crop in a particular year.
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble, distribution)?	No.
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	No.
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support.	Regional estimates of yield could be obtained from the June Survey data (see the “Historical Crop Rotation Data: June Survey data” observation table) – the current survey (2010) is a full census, and so this should provide an accurate estimate for the current year when this becomes available.
Will visualisation be necessary? How might this be done?	Yes. We want to be able to visualise how the estimated yield changes over the simulated years for each crop.
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	No. If we start the simulation from the past and simulate to the present day, we may be able to use the June Survey data to validate this output – see above.

3.3 Questions the model chain would address

The land-use experts are interested in using the model chain to explore the possible cropping conditions in regions of England under current conditions. We would like to investigate the impact of propagating uncertainty through to the yields as they are currently ignored. In order to do this, we will need to elicit (potentially complex) distributions for the uncertain inputs and to add appropriate uncertainty to the observation components. Also, due to the computational overhead of the LandSFACTS model, we expect to employ stochastic emulation techniques to make uncertainty analysis feasible.

Another question of interest is which parts of the land capability drive the production of which crops. As the specification of the distributions for the model inputs will be difficult for a lot of the uncertain inputs, this may amount to scenario modelling. This particular use of the model chain is of most interest to policy makers as they could investigate climate impacts.

In terms of visualisation, the land-use experts like to produce maps of fields for different years and regions that they have analysed. We need to produce similar maps that show potential changes in cropping over the period simulated. We could also show how this

changes for different land capability scenarios. These graphics could be used to highlight the impact of changing land capability on cropping and subsequent agricultural output.

4 Probabilistic forecasting of air quality (WP6)

This section reviews the requirements for the air quality application domain.

Frequently used acronyms (in this section)

AQ: Air Quality

ECMWF: European Centre for Medium range Weather Forecasting

EPS: Ensemble Prediction System

GEMS: Global and regional Earth-system (Atmosphere) Monitoring using Satellite and in-situ data project

GRIB: GRIdded Binary

MACC: Monitoring Atmospheric Composition and Climate

TAPM: The Air Pollution Model

4.1 Overview of the Model Chain

The aim of the application is to provide probabilistic forecasts (up to 2 days) and nowcasts (current and coming 2-3 hours) for air quality in Oslo. To achieve this a model chain is required that delivers output from synoptic scale weather forecasts (ECMWF) to a mesoscale meteorological model (TAPM) which in turn provides urban scale meteorological fields as input to an urban and local scale air quality model (EPISODE). Additional input to EPISODE includes emissions and background concentrations. This chain will provide the 2 day forecast. In addition to the 2 day forecast a short term (3 hour) forecast will be delivered that will combine near real time observations with the modelled air quality (data assimilation).

To provide probabilistic forecasts ensemble methods will be employed where both input data from other models and input parameters are provided as ensembles. The major scientific aim of this WP is to develop and implement a probabilistic air quality forecast system. In regard to UncertWeb a number of project aims are also identified

- To provide a test case where a number of existing and soon to be available data sources are implemented using UncertWeb services
- To provide a test case where ensemble methods are dealt with in UncertWeb
- To provide a test case where a number of model components are chained using UncertWeb services
- To provide a test case for processing and representation of uncertainty based on ensembles

An overview of the model chain is provided concisely:

Input data to the model chain:

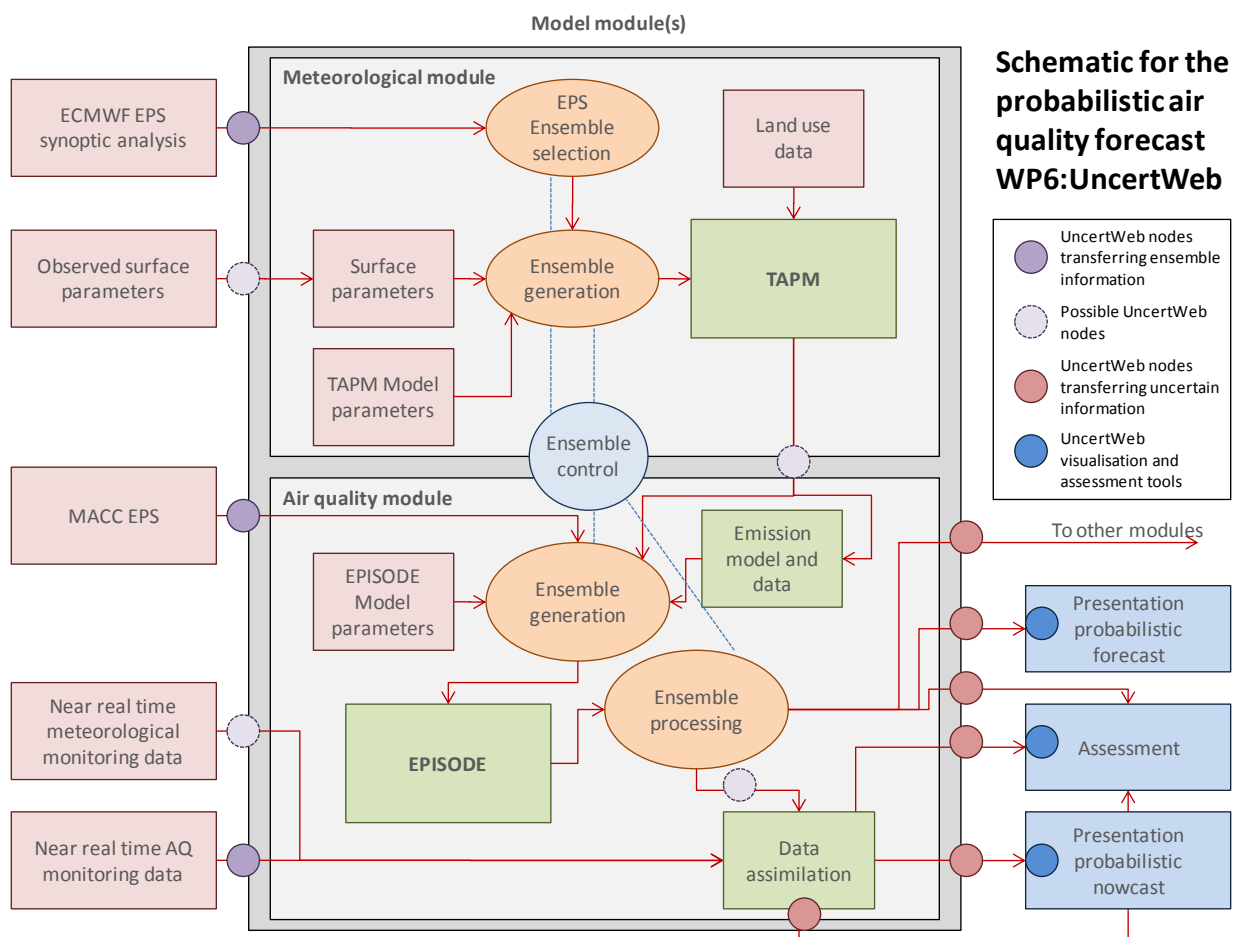
1. ECMWF synoptic scale ensemble forecasts. These are available real time (for money) or in archived form (free). The ensemble forecasts represent different outcomes given perturbations of the initial states for the model. These are available from ECMWF in GRIB format
2. Boundary conditions/background concentrations provided by GEMS models. This is an ensemble (10) of regional scale air quality models predicting air quality in Europe. Data is accessible through an ftp site and are in GRIB2 format

3. Emissions data for Oslo and environment. Uncertainty in these can be provided as an ensemble of emissions or may also be communicated as extra information, e.g. pdfs for each ensemble member
4. A list of uncertain model and surface parameters that will form part of the input ensemble data set.

The input data to the model chain are essentially uncorrelated, with the exception of the regional air quality and the synoptic meteorology. Regional scale models will use the same or very similar meteorology to the base synoptic meteorological field.

Currently in the flow chart the urban scale meteorological model (TAPM) and the urban scale air quality model (EPISODE) is represented by the ‘Air quality forecast model’. In fact they are two separate models, one providing input for the other. This leads to the question: Do we try to link the TAPM output with the EPISODE input using UncertWeb services (i.e. separate model components) or treat these as a single model component.

A more detailed flow chart is available below containing this possibility



4.2 Components in the Model Chain

These will compose of models and data sources.

4.2.1 Model components

General information about the model component

Model name (make this unique)	TAPM-EPISODE
Model creator (where does it come from?)	CMAR-CSIRO and NILU
Model licence (who can use it? Is it open source?)	TAPM: Licensed executable model. No code available but developers are subcontractors EPISODE: Licensed model from NILU, code available to consortium
Model requirements (operating system, hardware)	Both are PC windows based systems
Model computation (typical time to evaluate given a set of typical inputs – may require some commentary)	TAPM: Typical one day simulation will take around 5 -20 minutes, dependent on resolution. Tradeoffs usually required in terms of resolution. These times are based on fairly normal modern PC's with single processors. If we consider 100 ensemble members then the total CPU time will be around 33 hours. Parallel processors will be used. 8 processors will allow for run times of around 4 hours, which is acceptable for a 2 day forecast. EPISODE: For Oslo, one days simulation of air quality including all sources will take around 10 minutes
Model description (<i>overview</i> of what the model does, how it works – this will need to be brief)	TAPM is used as a meteorological model. It takes synoptic analysis at 125 -50 km resolution and nests these down (3-4 nestings) to 1 km resolution in the area of interest. EPISODE is an air quality model that takes in gridded meteorological data (in this case from TAPM) and combines these with a range of emissions sources and models to produce concentrations in fields and at 'receptor' points. In this application the 2 models are used together. Hourly meteorological fields from TAPM are given to EPISODE and hourly mean concentrations are calculated.
Relevant references (papers or other resources, including web links, that describe the model)	TAPM www.cmar.csiro.au/research/tapm/ (many references available here) EPISODE www.nilu.no/AQM/ The urban air dispersion model EPISODE applied in AirQUIS2003. Technical description. (http://www.nilu.no/index.cfm?ac=publications&folder_id=4309&publication_id=4653&view=rep)
Website or link for downloading the model, if	TAPM www.cmar.csiro.au/research/tapm/

available.	
Additional comments	Question remains on how to deal with this model system. Should these two models be connected via UncertWeb services or treated as a single model component? In principle they are a chain and so would fit with the UncertWeb principle. This would also allow for modulating these two often separate model types. In the rest of this description it will be assumed that these two models are linked through UncertWeb. However, alternative direct links are also possibly and will also likely be implemented

Summary of component model inputs

TAPM

Input name	Short description	Uncertain?	Importance rank
TAPM: ECMWF ENSEMBLE (Synoptic meteorology)	<p>Large scale synoptic meteorology as boundary conditions for TAPM. These include wind, temperature, water vapour and pressure as 3D fields. Varying uncertainty, dependent on synoptic situation. Can be quite certain or very uncertain for a 2 day forecast. This is the major input to the TAPM model. Other inputs are predefined but relevant for uncertainty assessment. These are supplied as 51 ensembles. The ensembles are generated by perturbing the initial conditions.</p> <p>As alternative to the ensemble from ECMWF it is worth noting that TAPM can also make use of the NCEP forecasts. It is not known yet how much variability the ECMWF ensembles introduce on the short term forecasts, after they have been downscaled by TAPM, so an alternative could be to access two different models (ECMWF and NCEP) and use these as input to the ensemble.</p>	Yes, but not certain by how much. To be handled by UncertWeb	2- 7
TAPM: LAND USE (Land use data)	<p>Obtained from a global land use database .This database can be manually altered for the situation (2D field). Land use data is used in the model to supply surface parameters for the model typical of that land use type. This field is static so any errors in it will lead to bias in the model.</p> <p>This will not be treated as uncertain input in UncertWeb and will be part of the model module</p>	Yes Not to be handled by UncertWeb	7
TAPM: SURFACE PARAMETERS (Initial surface parameters)	There are a number of surface parameters, often linked to the land use data that need to be supplied as initial conditions to the model before a calculation can be made. These include: Soil moisture content, soil temperature, sea surface temperature and	Yes. Observation as part of this to be handled by UncertWeb	2 - 7

	<p>snow and ice cover. When no other information is available then TAPM supplies default values and climatological values. Some of these are uncertain, e.g. soil moisture content and others are more certain, e.g. sea surface temperature. Some of these parameters may also be externally supplied from local measurements or information, e.g. sea surface temperature and snow and ice cover. Some sources of such information can be found at http://retro.met.no/kyst_og_hav/index.html .</p>		
TAPM: MODEL PARAMATERS (Model parameterisations and parameters)	<p>A range of choices is offered by the model for internal parameterisations of particular processes, e.g. non-hydrostatic pressure solver, turbulence scheme, precipitation scheme. The model may or may not be sensitive to these choices, dependent on the situation. Generally not very sensitive from my experience. Some constants used in the parameterisations are quite uncertain. These are generally not available to the normal user to vary but in the version used in UncertWeb a number of these parameters will be available.</p> <p>This will be part of the model module and will not require UncertWeb transfer, however if UncertWeb is to keep track of the ensembles then this must be connected somehow.</p>	Yes Not to be handled by UncertWeb	5 -7

EPISODE

Input name	Short description	Uncertain?	Importance rank
EPISODE: REGIONAL BACKGROUND (Regional scale air quality ensembles)	<p>A number of regional scale models (7-10) produce regional scale AQ forecasts for Europe (see http://www.gmes-atmosphere.eu/services/raq/). These are delivered daily to an ftp site (ftp://gems@data-portal.ecmwf.int) in GRIB2 format. Each model provides European wide 3 day forecasts. In addition to each model also an 'Ensemble' file should be available. This may contain for instance medians, means and standard deviations of the ensemble at different sites. This service is not yet completely established and at the moment only the individual model calculations are available. In Oslo there is not usually a large contribution from the regional background</p>	Yes. To be handled by UncertWeb	5

	but it is definitely important elsewhere.		
EPISODE: EMISSIONS (Emission data)	Emission data is required for the AQ forecast. Some of these data are quite certain whilst others are not. Some emission data is linked to the uncertain (ensemble) meteorological forecasts coming from TAPM.	Yes Not to be handled by UncertWeb	2
EPISODE: MODEL PARAMETERS (Model parameters)	As in TAPM there are a number of model parameters and parameterisations that are uncertain. These can be specified and used in the ensemble	Yes Not to be handled by UncertWeb	2-5

DATA ASSIMILATION

Input name	Short description	Uncertain?	Importance rank
EPISODE: ENSEMBLE OUTPUT	Results of the ensemble output from EPISODE. This will be in the form of 2D concentration fields every hour and as point calculations at the monitoring sites.	Yes. Possibly to be handled by UncertWeb	2
AQ MONITORING:	Near real time air quality monitoring data	Yes	7
METEO: MONITORING	Near real time meteorological data	Yes	7

Description of uncertain model inputs

Input name (unique if possible – add model name e.g. model:input)	TAPM: ECMWF ENSEMBLE
Input role (parameter in model, initial condition, boundary condition)	Boundary condition for TAPM
Input type (continuous, categorical, binary etc)	51 x 3D fields with 9 components 51 x 2D fields of ~5 components (to be determined)
For spatial inputs what is the domain (extent of the input region)	(65° N, 0° E, 55° N, 20° E) approx 40 x 80 x 9 grid points per ensemble member (this may be reduced after testing)
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	0.25 x 0.25 degrees as gridded data with 9 vertical levels
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Grid cells with an interpolated central location value
For temporal inputs what is the domain (extent of the input time – at a time point,	6 hourly data representing the instantaneous hourly mean values at that time point. 2 days is the forecast period.

over an interval?)	
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	Time series of 2D and 3D fields at 6 hour intervals as defined in the GRIB formatting
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	1 hour
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	Ensembles are used. Perturbation occurs in the initial conditions and 51 ensemble members are produced. The variability represents the uncertainty. There is no doubt more information available on this but this is not known.
Is the marginal (at a single location or object) probability density function known for this input?	Probably a lot of analysis has been done on this but this information is not currently known.
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	No, but this may have been studied. Correlation between ensembles will generally be high.
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	No
Do you know, and can you specify typical ranges for this input?	Typical ranges are typical meteorological situations.
Input source (describe where this input might come from – link to a observation component if this makes sense)	This data is available from ECMWF in the form of archived data and in real time data. The data must be downloaded from the ECMWF ftp server before it can be accessed.

Input name (unique if possible – add model name e.g. model:input)	TAPM: SURFACE PARAMETERS
Input role (parameter in model, initial condition, boundary condition)	Set of initial conditions for the surface model used in TAPM
Input type (continuous, categorical, binary etc)	1D and 2D values for the domain
For spatial inputs what is the domain (extent of the input region)	The domain is the TAPM domain. From 500 – 40 km dependent on the nesting grid
For spatial inputs what is the	Gridded fields according to the TAPM nested grids

resolution (and how is this represented – grid, spectral, other)	
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Grid cells with an interpolated central location value
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	These parameters are initial but climatological sea surface data is provided monthly
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	These parameters are initial
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	These parameters are initial
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	When monitored data is used for the surface parameters this will be quite certain. Uncertainty comes to a large degree in the fact that each square can only have one land use type (mixed squares are not possible). This, and misrepresentation of the land use type, are the largest uncertainties and can be significant.
Is the marginal (at a single location or object) probability density function known for this input?	No
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	No
Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	No
Do you know, and can you specify typical ranges for this input?	Depends on the parameter of course, but yes ranges can be specified.
Input source (describe where this input might come from – link to a observation component if this makes sense)	The data used is in the form of an existing database, any changes to this will be based on observations. These would be limited to: <ol style="list-style-type: none"> 1. Sea surface temperature 2. Surface and deep soil temperature 3. Soil moisture content

	<p>4. Snow cover and depth</p> <p>5. Sea ice cover and depth</p>
--	--

Input name (unique if possible – add model name e.g. model:input)	EPISODE: REGIONAL BACKGROUND
Input role (parameter in model, initial condition, boundary condition)	Provides boundary conditions of concentrations for the EPISODE model. This will be provided as an ensemble of 7-10 model calculations
Input type (continuous, categorical, binary etc)	Continuous
For spatial inputs what is the domain (extent of the input region)	One value for the EPISODE domain will be used
For spatial inputs what is the resolution (and how is this represented – grid, spectral, other)	Gridded fields at 25 km resolution
For spatial inputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Grid cell mean
For temporal inputs what is the domain (extent of the input time – at a time point, over an interval?)	Available as 3 day forecasts
For temporal inputs what is the time resolution (and how is this represented – time series, spectral, other)	Hourly
For temporal inputs what is time support (the period that one value in that input represents, e.g. time average or instantaneous value)	Hourly mean
Is input uncertainty known and to what degree? How is it described (probability density function, ensemble (number of members), realisation, summary statistics, other)?	Uncertainty is represented by the ensemble variability. This aspect would, will, be analysed by the source of the data, i.e. MACC
Is the marginal (at a single location or object) probability density function known for this input?	Will be known at other sites but not for Oslo
Is the joint (across the field, if a field input, typically given by a correlation function) probability density function known for this input?	No

Is the joint (across multiple inputs) probability density function known for this input with respect to other inputs? Is this a potentially important factor?	No
Do you know, and can you specify typical ranges for this input?	Depends on the parameter of course, but yes ranges can be specified.
Input source (describe where this input might come from – link to a observation component if this makes sense)	The data is available from an ftp site and can be downloaded from there. Data format is GRIB2.

Summary of component model outputs (*this is for all outputs of the model whether we treat them as of interest or not; for the important outputs more detail needs to be provided in the tables that follow*)

Output name	Short description	Importance rank
TAPM: ENSEMBLE OUTPUT	Ensemble of hourly 3D meteorological fields generated by TAPM for the 2 day forecast period	1
EPISODE: ENSEMBLE OUTPUT	Ensemble of hourly 2D Air Quality fields and 1D receptor points generated by EPISODE for the 2 day forecast period	1
DATA ASSIMILATION: OUTPUT	Hourly 2D Air Quality fields and 1D receptor points generated by the data assimilation module for the following 3 hour period	1
ENSEMBLE ASSESSMENT OUTPUT	Refers to the module for assessing and interpreting the ensembles. It's output will be uncertainty information for further presentation	1
PROBABILISTIC PRESENTATION OUTPUT	Output of the presentation (text and images) of the ensemble forecast	1

Description of uncertain model output

Output name in form model:output	TAPM: ENSEMBLE OUTPUT
Output role (feeds into / is used for)	Input to the emissions model and EPISODE air quality model
Output type (continuous, categorical, binary etc)	Hourly 3D fields and 2D surface fields for 2 day forecast period.
For spatial outputs what is the domain (extent of the input region)	40 x 40 km
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	1 km grid
For spatial outputs what is	Hourly mean at central location

support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	
For temporal outputs what is the domain (extent of the input time – at a point, over a time interval?)	Hourly values for the 2 day forecast period
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	Hourly, 3D gridded fields for each hour
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	Represents hourly mean values
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble, distribution)?	No, unknown, represented by the ensemble. Some assessment have occurred. And may occur during the project
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	Some errors are known for other application areas but this would have to be assessed for this application. Corrections to the model are not included and so bias may be present in the meteorological fields
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support. Please provide a complete description of the observations as a data component	Meteorological data is available at 3 sites in Oslo for this application. These are not usually available in real time but the possibility exists. They may be used for assessment purposes posterior. The spatial representative of the measurements (support) is under some conditions quite good but under others may not be so (dependent on the meteorological conditions).
Will visualisation be necessary? How might this be done?	Visualisation of the output meteorological fields, particularly wind fields is required. There are a number of established methods for this, e.g. wind roses, and wind vectors, temporal plots of wind speed and wind direction and stability parameters.
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	The output of the model has been, and will be, validated against the available data. Wind vector difference magnitudes are used for this as well as comparisons of stability. Standard methods are available for this that produce a number of statistical parameters indicating the quality of the model performance.

Output name in form	EPISODE: ENSEMBLE OUTPUT
---------------------	--------------------------

model:output	
Output role (feeds into / is used for)	This feeds into the assessment, processing and visualisation of the forecast as well as into the data assimilation module
Output type (continuous, categorical, binary etc)	Hourly 2D fields and point values for 2 day forecast period.
For spatial outputs what is the domain (extent of the input region)	20 x 20 km
For spatial outputs what is the resolution (and how is this represented – grid, spectral, other)	1 km grid
For spatial outputs what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Hourly mean at central location
For temporal outputs what is the domain (extent of the input time – at a point, over a time interval?)	Hourly values for the 2 day forecast period
For temporal outputs what is the time resolution (and how is this represented – time series, spectral, other)	Hourly, 2D gridded fields and 1D point values for each hour
For temporal outputs what is time support (the period that one value in that output represents, e.g. time average or instantaneous value)	Represents instantaneous concentrations (based though on hourly mean emissions) so almost the same as hourly mean values
Is output uncertainty currently known and quantified and to what degree? How is it described (samples, ensemble, distribution)?	For this combination and ensemble this is not currently known. Many assessments have been carried out in the past and can be used as indicative.
Is the model error (model inadequacy, structural error, model discrepancy) known for this output? Is this included in the model itself?	Some errors are known for other application uses but this would have to be assessed for this application. Corrections to the model results are not implemented
Are observations of this output available? How do the observations link to output – what is the sensor model or observation operator and are there issues of differing support. Please provide a complete description of the observations as a data component	Monitoring data is available from +/- 7 sites in Oslo. Most are traffic sites meaning that they are next to roads and can only be compared to the sub-grid receptor output. Even so, wind direction plays an important role and we always look at concentrations on both sides of the road. Only 1 or two sites would be representative of the 1 km grids used in the model.

Will visualisation be necessary? How might this be done?	Visualisation of the output air quality fields will be required. This is usually done as gridded data or as time series data for individual points. Also the information can be aggregated into a single air quality index that confers information on the ‘general’ air quality in the whole city. See www.luftkvalitet.info/ on how this is currently done (Norwegian).
Is this output currently validated (in the sense of comparing to observations)? How is this currently done?	The output of the model is assessed for the different applications against monitoring data. Time series and statistical analysis is carried out. Reports are produced yearly on this (in Norwegian).

4.2.2 Observation components

These are the observations that might be used to determine some inputs within UncertWeb, and in particular for validation of outputs. Each application WP is likely to need several of these, and some are likely to be shared. They are separate from model components since they might exist as services outside the particular model chain.

Data set name (unique)	Meteorological observations
Data source (where does the data come from)	Two synoptic meteorological sites (available in real time) and one specialised meteorological site run by the Oslo commune.
Data availability (where can this be obtained from, access rights)	2 synoptic sites are displayed by the local met bureau (www.met.no , www.yr.no). These data are in principle available. The specialised site is not normally available on line but the data is transferred to NILU and stored as ascii files on an internal server
Data accessibility (is this available over the web, and how is this made available?)	Synoptic stations can be visualised on the web through the meteo web sites. Not clear of accessibility otherwise. The specialised site has the data available but not currently for the web.
Data type (continuous, categorical, binary etc)	Continuous time series of a number of meteorological parameters
For spatial observations what is the domain (extent of the input region)	Point sampling
For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)	Point sampling
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Measurements are representative of a region perhaps 100 m around the site under normal conditions but representative of a larger region under some meteorological conditions (e.g. strong winds). In general the meteorological information contained at the specialised site has been found to be generally representative for Oslo, given that there can be significant variation of the Oslo region.
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	Continuous measurements

For temporal observations what is the sampling frequency.	Sampling frequency is around 5 minutes but 1 hour averages are the norm.
For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	Hourly averages
Is uncertainty known and is this broken down into observation errors, processing errors, representativity? How is it described?	No assessment of the uncertainty of the meteorological measurements has been made. Standard information concerning the accuracy of the instruments is provided by the manufacturers. Spatial representativeness is likely the largest uncertainty source. Data is manually quality controlled at some point in time (months after the measurements)
Do you have confidence in being able to specify uncertainties? What issues can you foresee in specifying uncertainties?	Uncertainties are limited for these instruments (excluding representativeness). Malfunctions and freezing are the largest source of uncertainty and quality control of the data is required to assess this.

Data set name (unique)	Air quality observations
Data source (where does the data come from)	Oslo has up to 7 air quality stations available for a number of standard compounds. Stations are run by local authorities
Data availability (where can this be obtained from, access rights)	Data is freely available. Real time data storage and web presentation is already established. Database and web solution situated at NILU
Data accessibility (is this available over the web, and how is this made available?)	Available over the web
Data type (continuous, categorical, binary etc)	Continuous time series of a number of air quality parameters
For spatial observations what is the domain (extent of the input region)	Point sampling
For spatial observations what is the sampling resolution (and how is this represented – points, grid, spectral, other)	Point sampling
For spatial observations what is support (the area that one value in that input represents, e.g. for a grid is it the grid cell mean or the central location value)	Traffic stations are representative of a small region, e.g. 10 m from the road but representative of the along road distance. Traffic stations are sensitive to wind direction that can blow the pollution either towards or away from the station.
For temporal observations what is the domain (extent of the input time – at a point, over an interval?)	Continuous measurements
For temporal observations	Sampling frequency is around 15 minutes but 1 hour

what is the sampling frequency.	averages are the norm.
For temporal observations what is time support (the interval that one value in that input represents, e.g. is it a time average, or instantaneous value)	Hourly averages
Is uncertainty known and is this broken down into observation errors, processing errors, representativity? How is it described?	This will need to be determined. The uncertainty is different before and after calibration (carried out weekly) and quality control (carried out monthly). Real time data can be poor for NO ₂ before the weekly calibration. Particulate matter measurements have an hourly uncertainty of around 10-30%
Do you have confidence in being able to specify uncertainties? What issues can you foresee in specifying uncertainties?	This will need to be determined. Expert elicitation and review of methods is a likely source of information.

4.3 Questions the model chain would address

The following points reflect aspects that UncertWeb can contribute to the probabilistic AQ forecasting system:

1. There is a need to communicate the necessary information (metadata) relating to each ensemble member, as well of course as the data for each ensemble member, in order to keep track of the ensembles and analyse the results effectively at the end of the chain.
2. Presentation and assessment are two areas where UncertWeb can provide tools. These need some long discussions to decide what is feasible.
3. Once the final ensembles are generated a method for communicating the results further to other modules, either as an ensemble or as a probability density function, is required. This relates to WP8.
4. The ensemble data used is from 1 to 4 dimensional data and UncertWeb will need to deal with this sort of spatial data appropriately
5. Some form of aggregation will be required in order to communicate the predicted air quality in the most effective way. Often used now is the air quality index which is based on maximum concentrations aggregated spatially and over the different pollutants.

5 Summary

This report has presented the use-cases and scenarios which we will implement within the UncertWeb project. Together these form a crucial step in defining the requirements for the work within WP1-3 which will define the scope and boundary cases that the UncertWeb solutions will be able to address. Also, the requirements form the content for WP8, which focuses on the integration of two of the use cases.

Creating these descriptions of the model chains, and the components and data flows within the model chains emphasises the diversity of applications addressed within UncertWeb, but also shows the many common factors, in that several data sets, for example weather and climate information, are shared across all application areas.

The next steps involve a thorough analysis of the requirements to capture requirements related to representation of uncertainty (D1.1), requirements for discovery and chaining (D2.1) and requirements for usability and integration (D8.1).