

Spatial data quality and error propagation in spatio-temporal modelling, **in practice**

Edzer Pebesma



WESTFÄLISCHE
WILHELMUS-UNIVERSITÄT
MÜNSTER



ifgi

Institute for Geoinformatics
University of Münster

edzer.pebesma@uni-muenster.de

7th International Symposium on Spatial Data Quality
(ISSDQ 2011) Oct 12-14, 2011, Coimbra, PT

Spatial data quality and error propagation in spatio-temporal modelling, and how it works in practice

Edzer Pebesma, *with help from many others*



WESTFÄLISCHE
WILHELMUS-UNIVERSITÄT
MÜNSTER

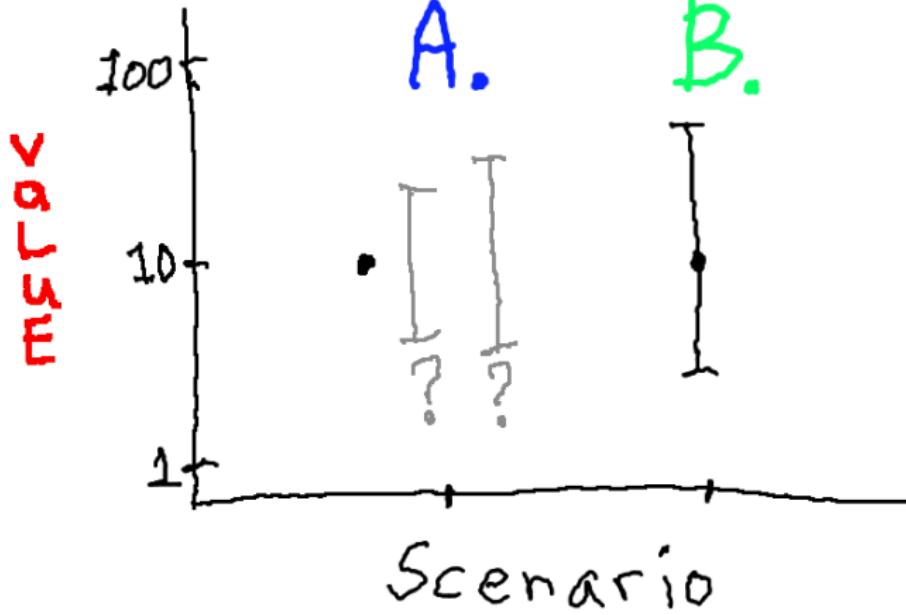


ifgi

Institute for Geoinformatics
University of Münster

edzer.pebesma@uni-muenster.de

7th International Symposium on Spatial Data Quality
(ISSDQ 2011) Oct 12-14, 2011, Coimbra, PT



How would the world look like...

How would the world look like...

How would the world look like when everyone used R?

How would you like to be remembered?

How would you like to be remembered?

+Tu Web Imagens Vídeos Mapas Notícias Gmail Mais ▾

Google académico [Pesquisa avançada do Google Académico](#)

Pesquisar a Web Pesquisar páginas em Português

Académico

Dica: [Pesquisa unicamente para resultados em Português \(Portugal\)](#). Pode indicar o seu idioma de pesquisa em [Preferências do Google Académico](#).

Gstat: a program for geostatistical modelling, prediction and simulation [\[PS\] de gstat.org](#)

EJ Pebesma... - Computers & Geosciences, 1998 - Elsevier
 ... The complete functionality of gstat is documented in its user's manual (**Pebesma**, 1997). ... notation:
 (1b). with $Z(s) = (Z(s_1), \dots, Z(s_n))'$, with known base functions $X = (x_0, x_1, \dots, x_p)$, $x_j = (x_{j1}, \dots, x_{jn})'$, with $\beta = (\beta_0, \dots, \beta_p)'$ the unknown parameter vector, with $e(s) = (e(s_1), \dots, e(s_n))'$
[Citado por 287 - Artigos relacionados - Todas as 10 versões](#)



Multivariable geostatistics in S: the gstat package* 1 [\[PDF\] de psu.edu](#)

EJ Pebesma - Computers & Geosciences, 2004 - Elsevier
 ... Permissions & Reprints. Multivariable geostatistics in S: the gstat package *1. Edzer J. **Pebesma**
 Corresponding Author Contact Information , E-mail The Corresponding Author. Department of
 Physical Geography, Utrecht University, PO Box 80.115, 3508 TC, Utrecht, Netherlands ...
[Citado por 264 - Artigos relacionados - Todas as 14 versões](#)

Applied spatial data analysis with R [\[PDF\] de ethz.ch](#)

RS Bivand, EJ Pebesma, V Gómez-Rubio... - 2008 - stat.math.ethz.ch
 ... book{Bivand-etal-2008, AUTHOR=Bivand, RS, **Pebesma**, EJ, Gómez-Rubio, V ... n sample size
 si location (si $\in G$) y(si) observation recorded at location si Y (si) random variable used to model
 y(si) E(·) expectation Var [·] variance Cov [·, ·] covariance C(·) covariance function ...
[Citado por 184 - Artigos relacionados - Ver em HTML - Todas as 8 versões](#)

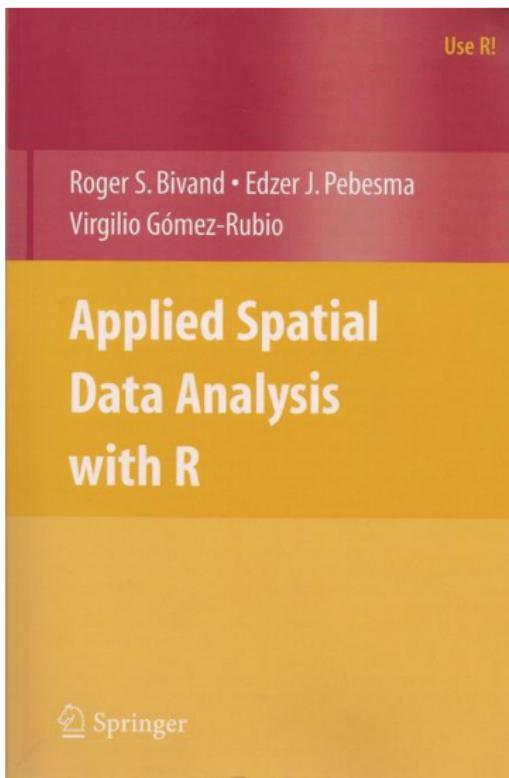
Spatial aggregation and soil process modelling

..., EJ Pebesma - Geoderma, 1999 - Elsevier
 ... PDF (894 K); Export citation; E-mail article; Highlight keywords on: ... Corresponding author.
 Tel.: +31-20-525-7448; Fax: +31-20-525-7431; E-mail: gbmheuvelink@frw.uva.nl.
 Geoderma Volume 89, Issues 1-2, April 1999, Pages 47-65. ...
[Citado por 101 - Artigos relacionados - Todas as 11 versões](#)

Latin hypercube sampling of Gaussian random fields

EJ Pebesma... - Technometrics, 1999 - JSTOR

R-spatial



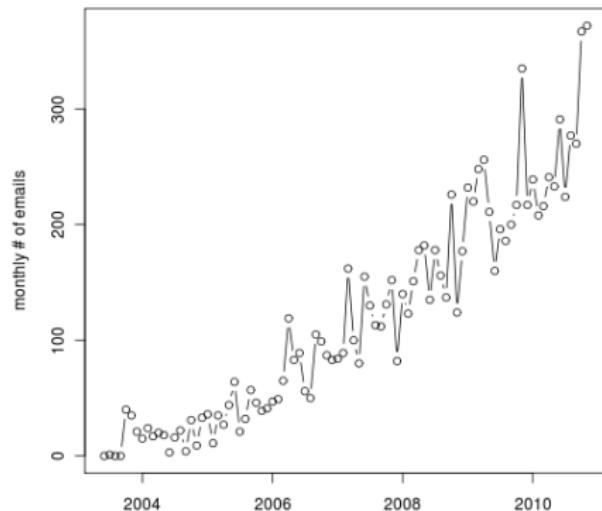
Use R!

Roger S. Bivand • Edzer J. Pebesma
Virgilio Gómez-Rubio

Applied Spatial
Data Analysis
with R

Springer

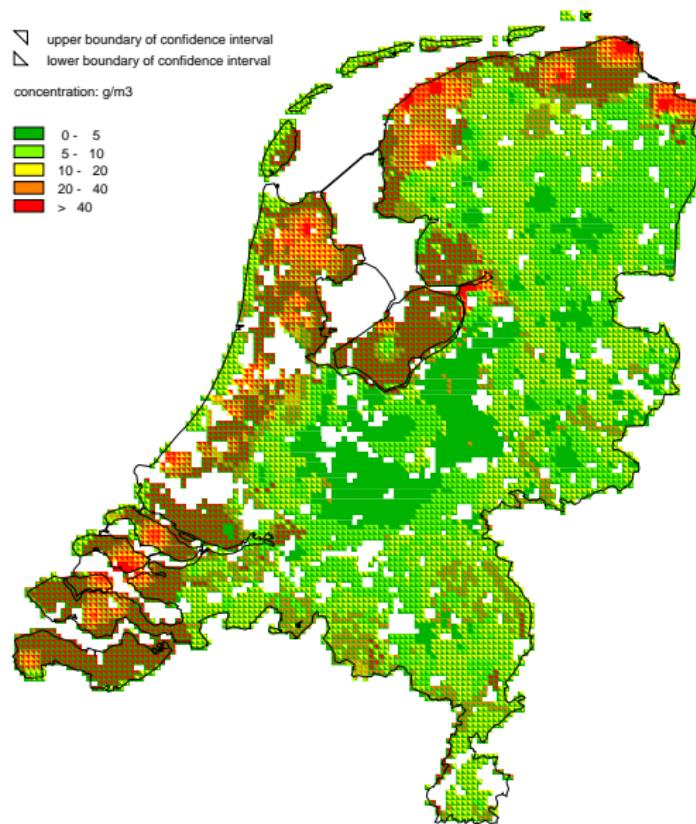
r-sig-geo email list traffic



Outline

- Raising awareness of Spatial Data Quality
- Two recent use cases
- Error propagation in practice: UncertWeb
 - Data Models
 - Tools: UTS, UPS, WPS4R, ...
 - Visualisation
- The future I: bottom-up
- The future II: top-down

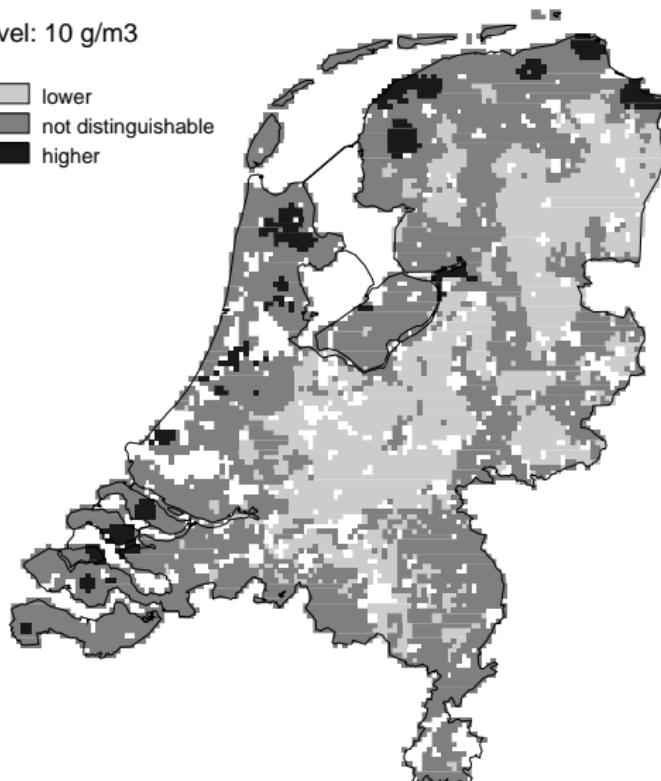
K, groundwater, 5-17 m depth



K, groundwater, 5-17 m depth

level: 10 g/m³

- [light gray] lower
- [medium gray] not distinguishable
- [black] higher



K, groundwater, 5-17 m depth

level: 5 g/m3

- lower
- not distinguishable
- higher



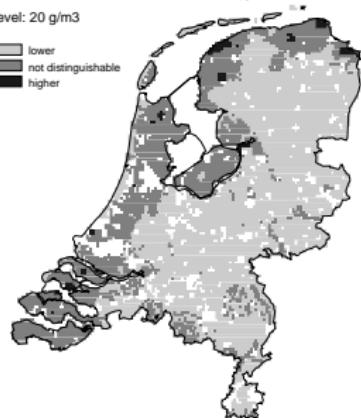
level: 10 g/m3

- lower
- not distinguishable
- higher



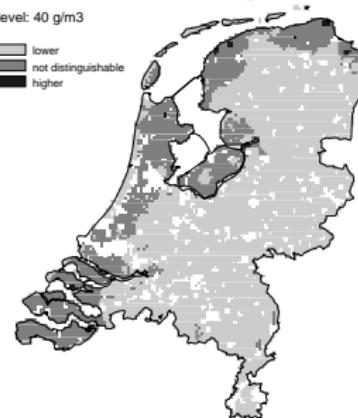
level: 20 g/m3

- lower
- not distinguishable
- higher



level: 40 g/m3

- lower
- not distinguishable
- higher



International Journal of Geographical Information Science
Vol. 21, No. 5, May 2007, 515–527



Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: an air quality example

EDZER J. PEBESMA[†], KOR DE JONG[†] and DAVID BRIGGS[‡]

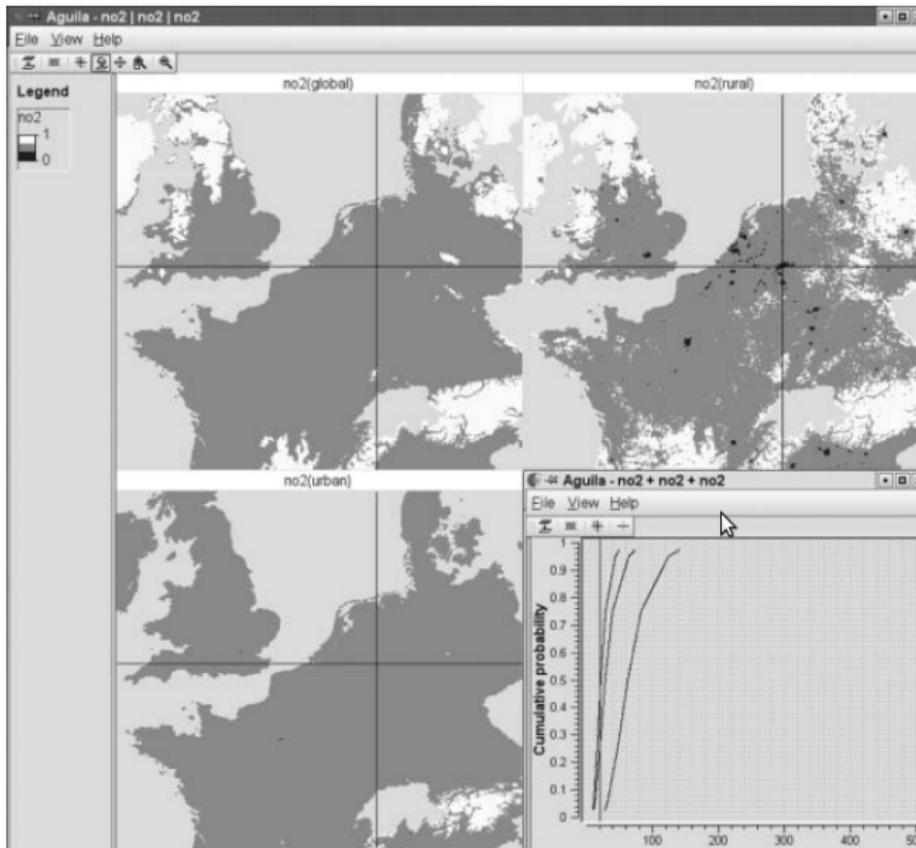
[†]Faculty of Geosciences, Utrecht University, The Netherlands

[‡]Imperial College, London, UK

(Received 23 November 2005; in final form 13 October 2006)

This paper introduces a method for visually exploring spatio-temporal data or predictions that come as probability density functions, e.g. output of statistical models or Monte Carlo simulations, under different scenarios. For a given moment in time, we can explore the probability dimension by looking at maps with cumulative or exceedance probability while varying the attribute level that is exceeded, or by looking at maps with quantiles while varying the probability

Same idea, but interactive

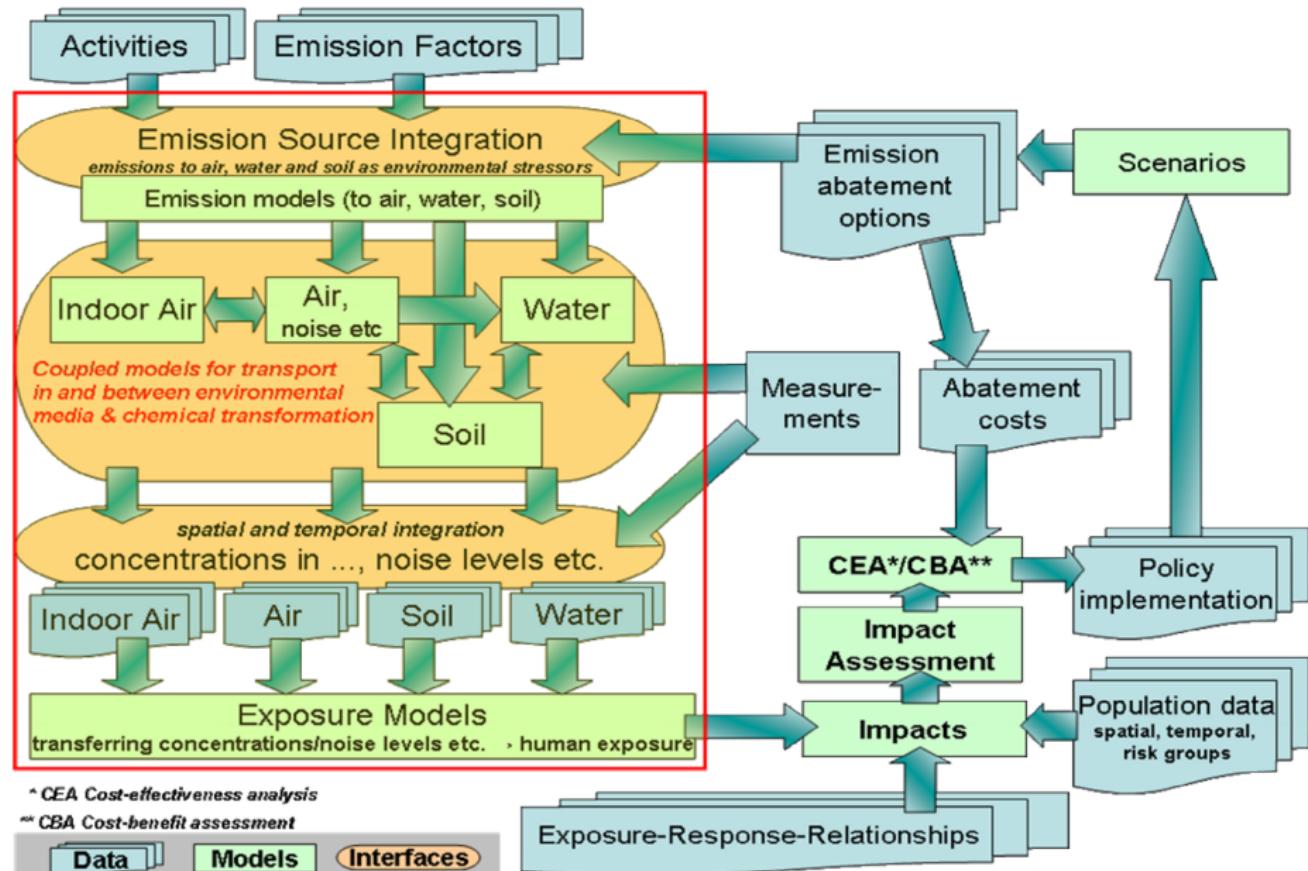


Two recent use cases

Uncertainty assessment of N₂O inventories

*Explorations at different spatial and temporal scales for the
Dutch fen meadow landscape*

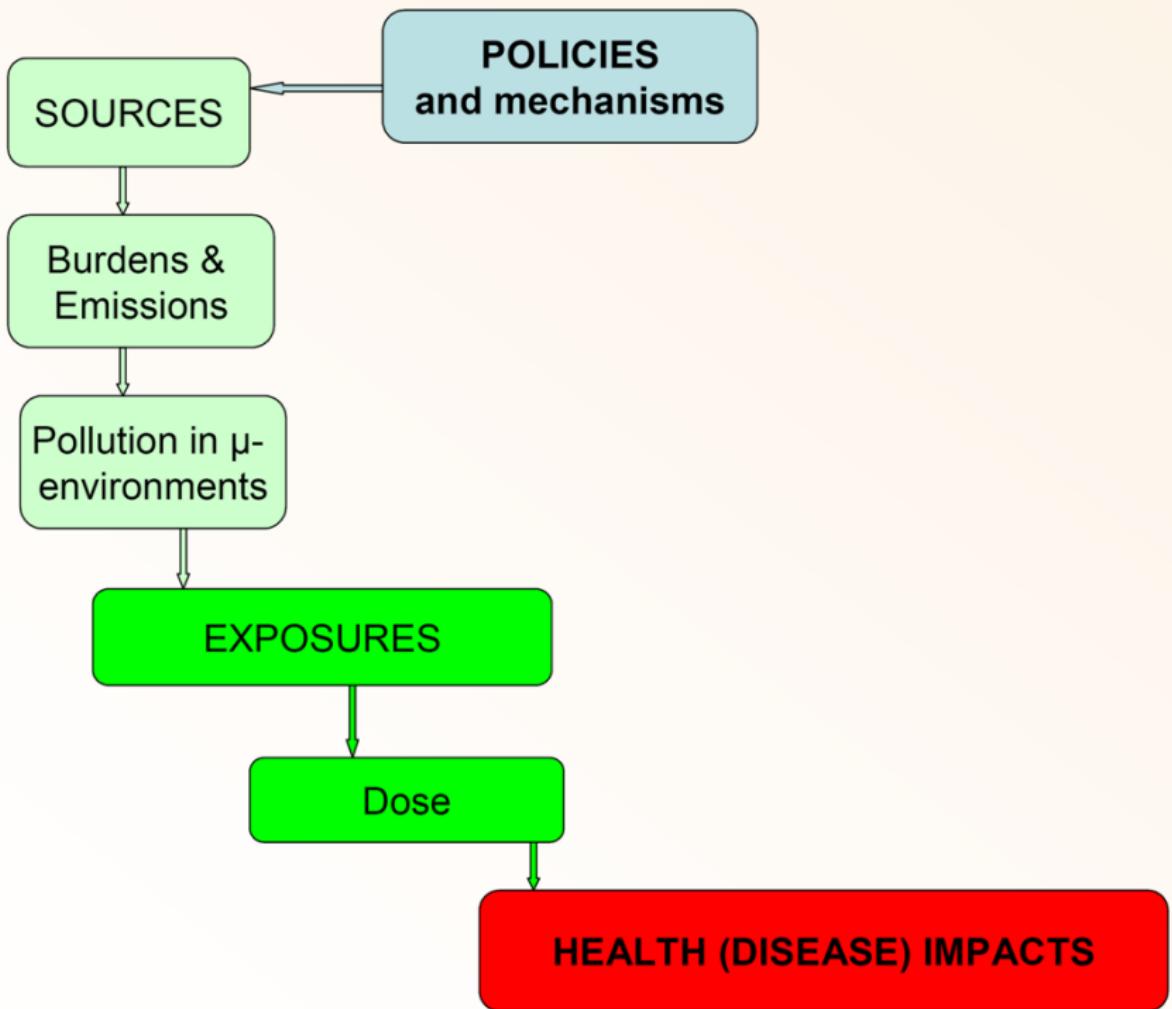
Linda Nol



* CEA Cost-effectiveness analysis

** CBA Cost-benefit assessment

Exposure-Response-Relationships



Why can't this be simple?

We should not underestimate the problem (dimensionality, inference) of the multivariate distributions we deal with.

Why can't this be simple?

We should not underestimate the problem (dimensionality, inference) of the multivariate distributions we deal with.

What can we do:

- wrap existing models in MC loops
- change to statistical models, hierarchical models
- compute with random variable primitives

The model web

The Model Web (GEOSS AR-09-02.d task) has the goal to *develop a dynamic modelling infrastructure (Model Web) [...] composed of loosely coupled models that interact via Web services, and are independently developed, managed, and operated.*

The model web – II

The principle is that models are exposed on the Web and can be **discovered**, **combined** into complex workflows and **executed** over a distributed architecture.

The model web – II

The principle is that models are exposed on the Web and can be **discovered**, **combined** into complex workflows and **executed** over a distributed architecture. Such a system provides tremendous opportunities to enhance scientific modelling by:

- improving the integration of different models to address practical questions;
- increasing the reproducibility and transparency of research by providing clear and repeatable provenance information for modelling outputs;
- allowing more flexible deployment, for example in cloud architectures;
- facilitating the discovery and reuse of model components and code.

The model web – II

The principle is that models are exposed on the Web and can be **discovered**, **combined** into complex workflows and **executed** over a distributed architecture. Such a system provides tremendous opportunities to enhance scientific modelling by:

- improving the integration of different models to address practical questions;
- increasing the reproducibility and transparency of research by providing clear and repeatable provenance information for modelling outputs;
- allowing more flexible deployment, for example in cloud architectures;
- facilitating the discovery and reuse of model components and code.

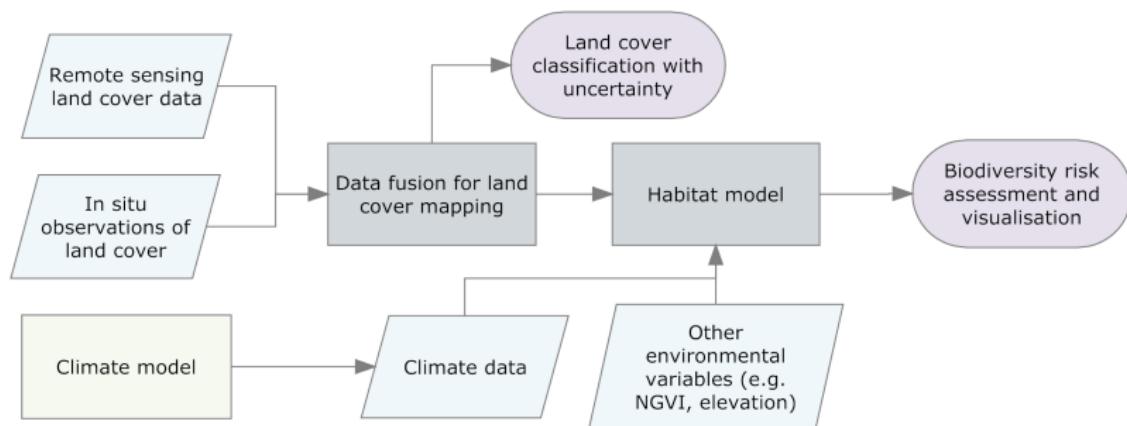
sensor web ⊂ model web

Error propagation in practice: UncertWeb

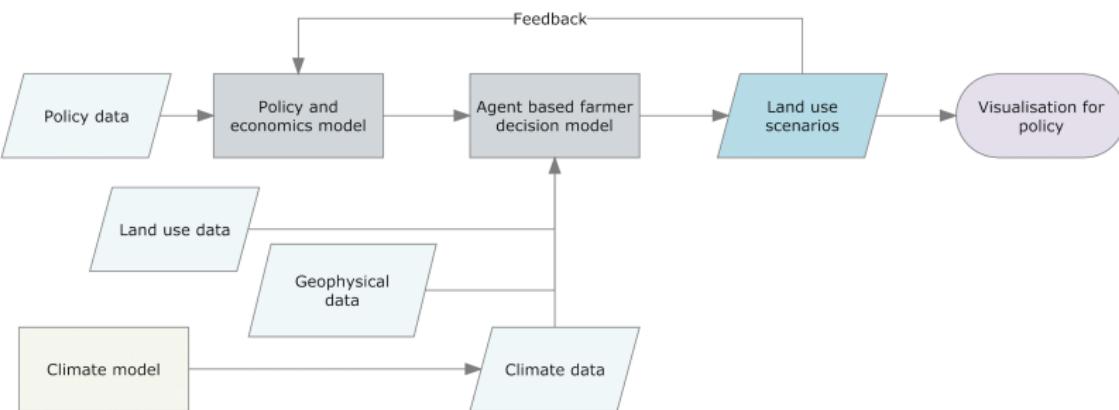
UncertWeb: the uncertainty-enabled model web

- UncertWeb will *uncertainty-enable* the Model Web
- Problem: the Model Web is still a vision:
 - its components (open, interoperable data and model services) do not yet exist (OGC? W3C?)
 - interoperable standards typically try to solve *everything*, making full support difficult (profiling).
- four case studies: biodiversity, land use change, air quality, human activity
- integration of 3+4
- data models: O&M-U, netCDF-U, UncertML
- orchestration, brokerage, mediation: CaaS (CNR)
- tools: elicitor, aggregator, UTS, UPS, visualizer

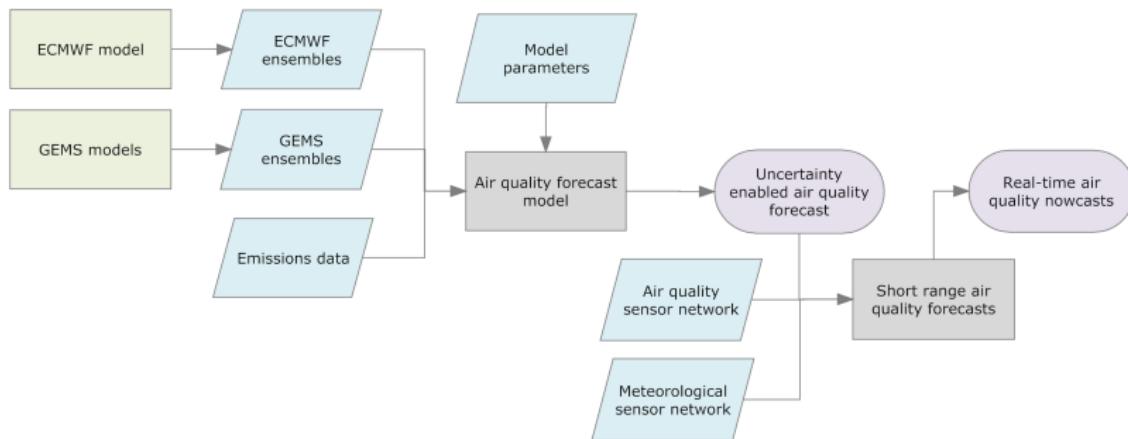
case: biodiversity



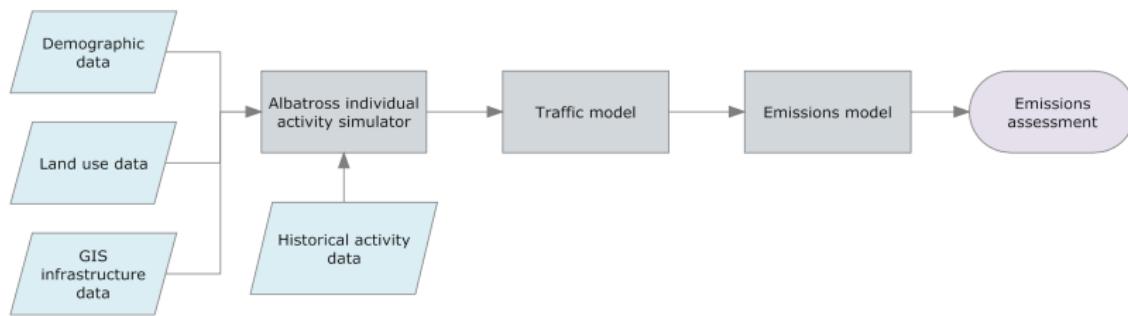
case: land use change



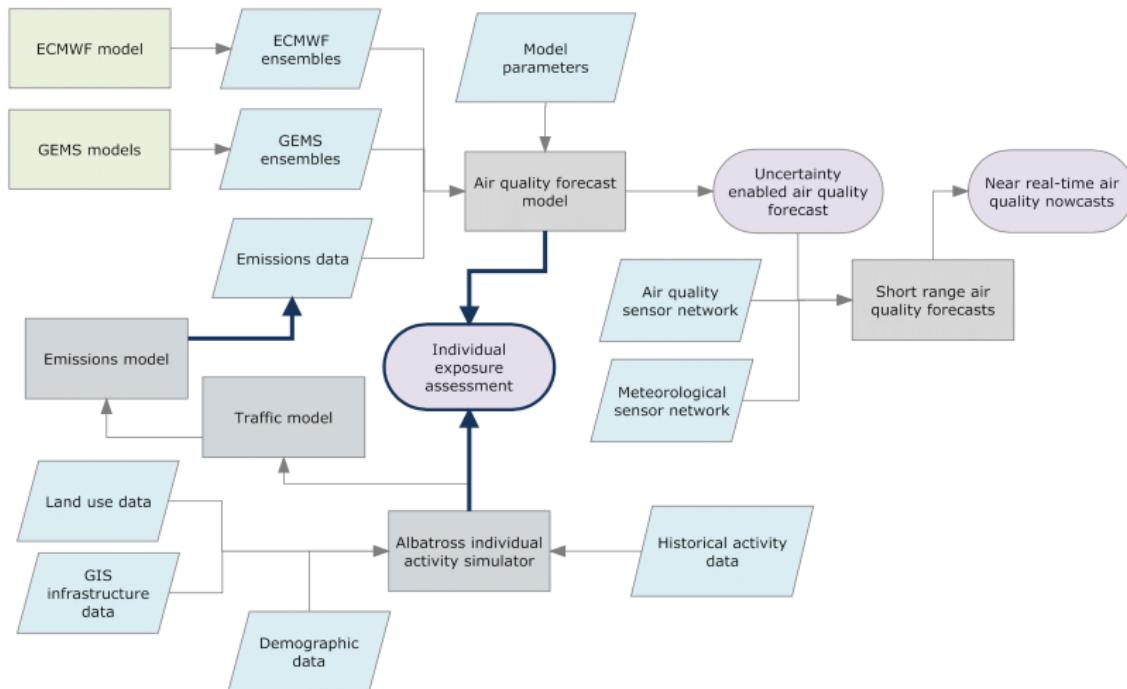
case: air quality



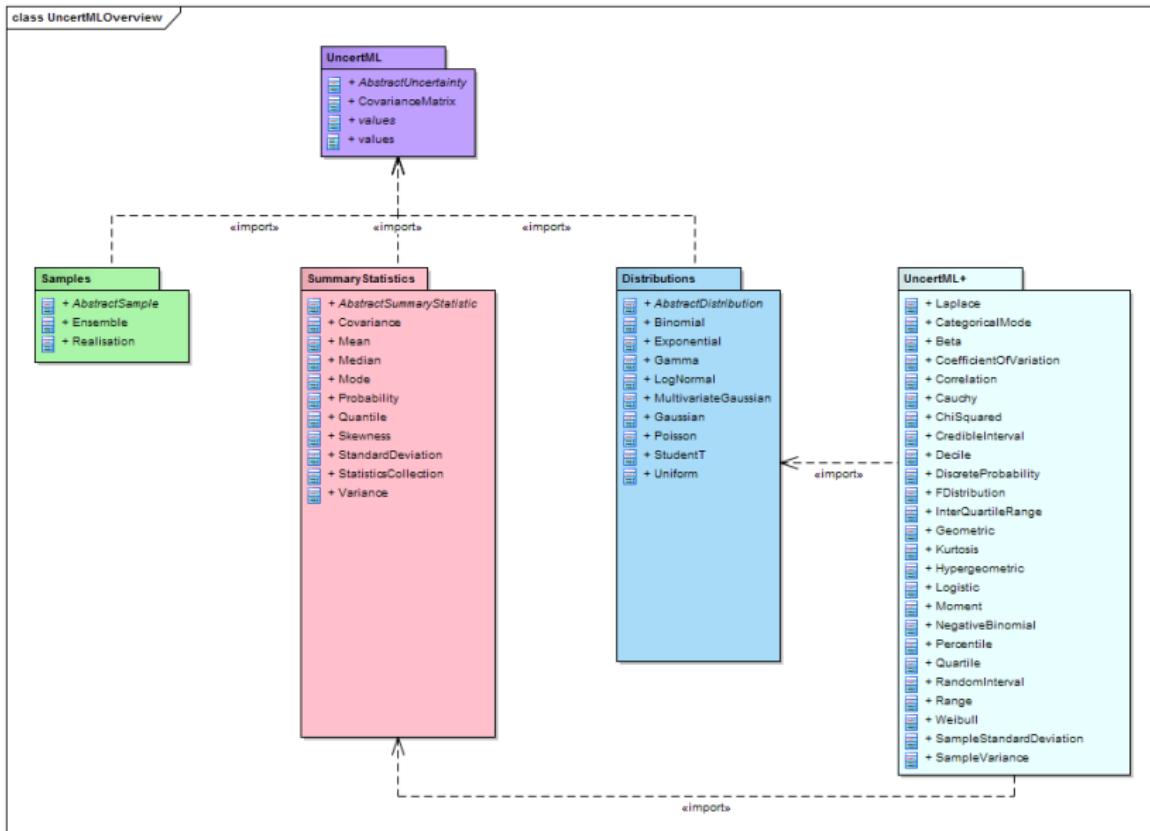
case: human activity



case: integrate air quality with human activity



Data models: UncertML



Data models: UncertML.org

The screenshot shows the homepage of UncertML.org. At the top, there is a navigation bar with links to "About UncertML", "API documentation", and "User guide". Below the navigation bar, there are three small images of heatmaps or spatial data. To the right of these images, the text "describing and exchanging uncertainty..." is displayed. On the left side of the page, there is a large blue sidebar. On the right side, there is a "Download" section with options for XML Schema, Java API, and Bundle. A note states that the bundle contains all UncertML dependencies. A "download" button is present, and the current release is v2.

UncertML.org

describing and exchanging uncertainty...

About UncertML API documentation User guide

Uncertainty Markup Language: UncertML

UncertML is a **conceptual model** and **XML encoding** designed for **encapsulating probabilistic uncertainties**. This website contains all the information you will need to start using UncertML to quantify and exchange your data uncertainties. For a detailed description of the models and XML schema, including use case examples, look no further than the [dictionary and user guide](#). Below are a series of questions and answers to give you a brief insight to UncertML.

[What is it?](#)

Download

- XML Schema (44 KB)
- Java API (4,254 KB)
- Bundle (3,478 KB)

The bundle contains all UncertML dependencies.

[download](#)

Current release: v2

Data models: O&M-U (UncertML)

Altova XMLSpy - [ObsCol_Point_TimeInstant_Double.xml]

File Edit Project XML DTD/Schemas Schema design XSL/XQuery Authentic DB Convert View Browser Tools Window Help

Project D:\M\GIN\OGC_material\SCHEMAS D:\M\GIN\OGC_SWGs\SOS\Branch D:\M\GIN\OGC_material\Vd_branch D:\M\GIN\Projekte\UncertWeb\WF D:\M\GIN\OGC\SWGs\ISO5\Spec I D:\M\GIN\Projekte\UncertWeb\WF D:\M\GIN\Projekte\UncertWeb\WF Profiles GML ISO_DQ OM examples DiscreteNumericObserver DiscreteNumericObserver Measurement.json MeasurementCollector ObsCol_Measurement ObsCol_Point_Timeinstant ObsCol_UncertaintyOb ObsLineString_Timelr Obs_Point_Timeinstant Obs_Polygon_Timeinstant Obs_Polygon_Timeinstant ReferenceObservation ReferenceObservation TextObservation json TextObservationCollection UncertaintyObservation UncertaintyObservation generateXBeans bat installMavenJar.bat UncertWeb_OM.xsd obUWProfiles jar org Sampling Catalog.xsd CustomCatalog.xml CustomCatalog_remote.xml deployMavenJar.bat generateXBeans.bat installMavenJar.bat obUWProfiles.jar Software

Elements Attributes Entities

```

<?xml version="1.0" encoding="UTF-8"?>
<?om:OM MeasurementCollection xmlns:om="http://www.opengis.net/2.0" xmlns:xs="http://www.w3.org/2001/XMLSchema-instance" xmlns:sams="http://www.opengis.net/samplingSpatial/2.0" xmlns:sf="http://www.opengis.net/sampling/2.0" xmlns:un="http://www.uncertml.org/2.0" xmlns:gmd="http://www.isotc211.org/2005/gmd" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:gml="http://www.opengis.net/gml/3.2" xsi:schemaLocation="http://www.opengis.net/om/2.0 http://mars.uni-muenster.de/uncertweb/schema/profiles/om/UncertWeb_OM.xsd">

  <!-- This observation depicts an example with
       -point as sampling location
       -time period as phenomenon time
       -result is a measure value (double with units of measure information)
       -no uncertainty information available -->

  <?om:OM_Measurement gml:id="obsTest1"?>
    <?om:phenomenonTime?>
    <?om:resultTime xlink:href="#of1t"?>
      <?om:procedure xlink:href="http://www.example.org/register/process/scales34.xml"?>
      <?om:observedProperty xlink:href="um:ogc:def:phenomenon:OGC_temperature"?>
      <?om:featureOfInterest?>
      <?om:resultQuality?>
        <?gmd:DQ_QualitativeAttributeAccuracy?>
          <?gmd:result?>
            <?gmd:DQ_UncertaintyResult?>
              <?gmd:valueUnit?>
                <?gmd:UnitDefinition gml:id="unit1"?>
                  <?gml:identifier codeSpace="http://unitsofmeasure.org"?>degC</?gml:identifier>
                  <?gml:UnitDefinition?>
                <?gmd:valueUnit?>
                <?gmd:value?>
                  <?un:NormalDistribution?>
                    <?un:mean>0</?un:mean>
                    <?un:variance>0.05</?un:variance>
                  <?un:NormalDistribution?>
                <?gmd:value?>
              <?gmd:DQ_UncertaintyResult?>
            <?gmd:result?>
            <?gmd:DQ_QualitativeAttributeAccuracy?>
          <?om:resultQuality?>
        <?om:result uom="degC">36</?om:result>
      <?om:result uom="degC">36</?om:result>
    <?om:result uom="degC">36</?om:result>
  <?om:OM_Measurement?>

```

Text Grid Schema/WSDL Authentic Browser

reservation.json UncertaintyObservationCollection.json Measurement.json ObsCol_Point_TimeInstant_Double.xml UncertWeb_DQ.xsd

Data models: NetCDF-U (UncertML)

The screenshot shows the 'Datasets Browser' application window. On the left, a tree view displays dataset variables:

Name	Long Name	Type
biotemperature_gnb	biotemperature_normalD...	Local File
biotemperature_mean		[lon][lat]
biotemperature_variance		[lon][lat]

On the right, the 'Dataset/Variable CDL Info' pane shows the CF-1.5 compliant CDL code:

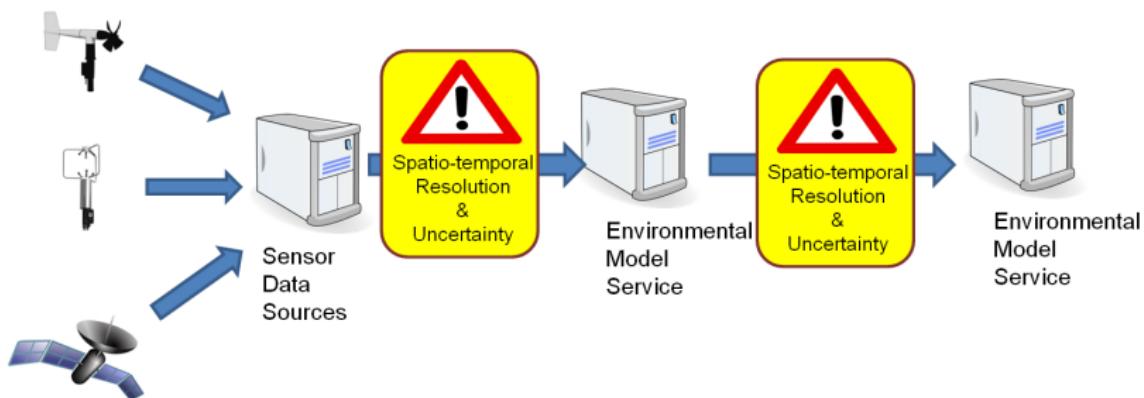
```
dimensions:  
  lon = 240;  
  lat = 163;  
variables:  
  double biotemperature_mean(lat=163, lon=240);  
    :missing_value = -999.0; // double  
    :ref = "http://www.uncertml.org/distributions/normal#mean";  
  double biotemperature_variance(lat=163, lon=240);  
    :missing_value = -999.0; // double  
    :ref = "http://www.uncertml.org/distributions/normal#variance";  
  double biotemperature;  
    :units = "degC";  
    :ancillary_variables = "biotemperature_mean biotemperature_variance";  
    :ref = "http://www.uncertml.org/distributions/normal";  
    :shape = "lat lon";  
  double lon(lon=240);  
    :long_name = "longitude";  
    :units = "degrees_east";  
    :_CoordinateAxisType = "Lon";  
  double lat(lat=163);  
    :long_name = "latitude";  
    :units = "degrees_north";  
    :_CoordinateAxisType = "Lat";  
  
:Conventions = "CF-1.5 UN-1.0";  
:primary_variables = "biotemperature";  
}
```

Tools: elicitor

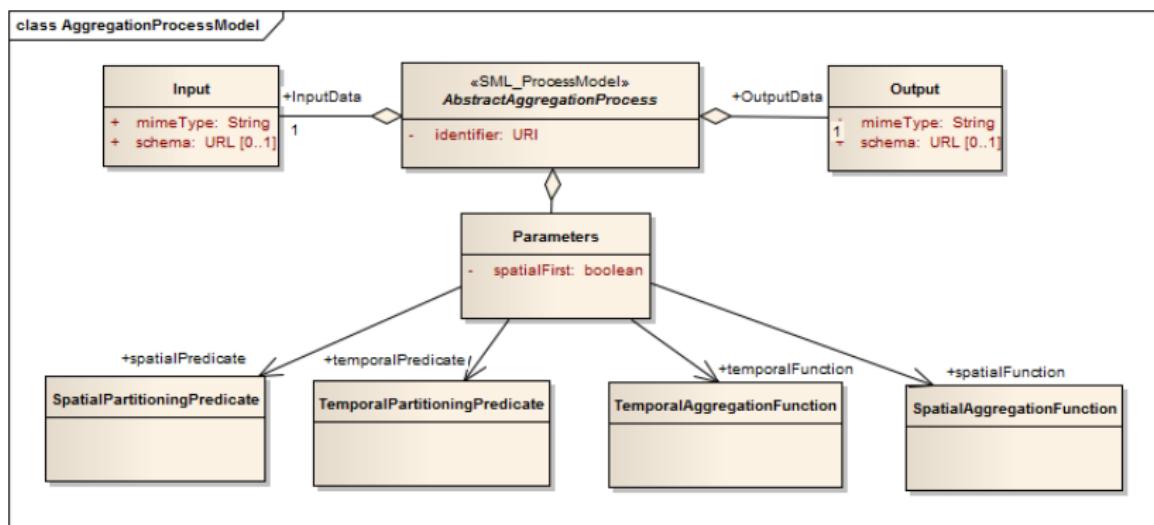
web-based expert elicitation tools to infer uncertainties from experts:

- univariate, combining different experts (Ast, Williams)
- spatially correlated variables (variogram elicitor, Heuvelink/Trong)

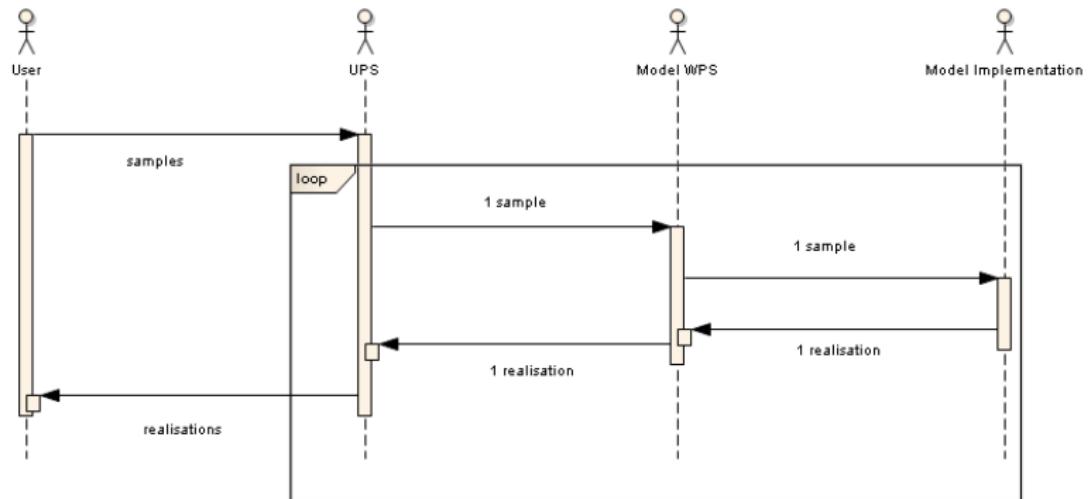
Tools: spatio-temporal aggregation service (STAS)



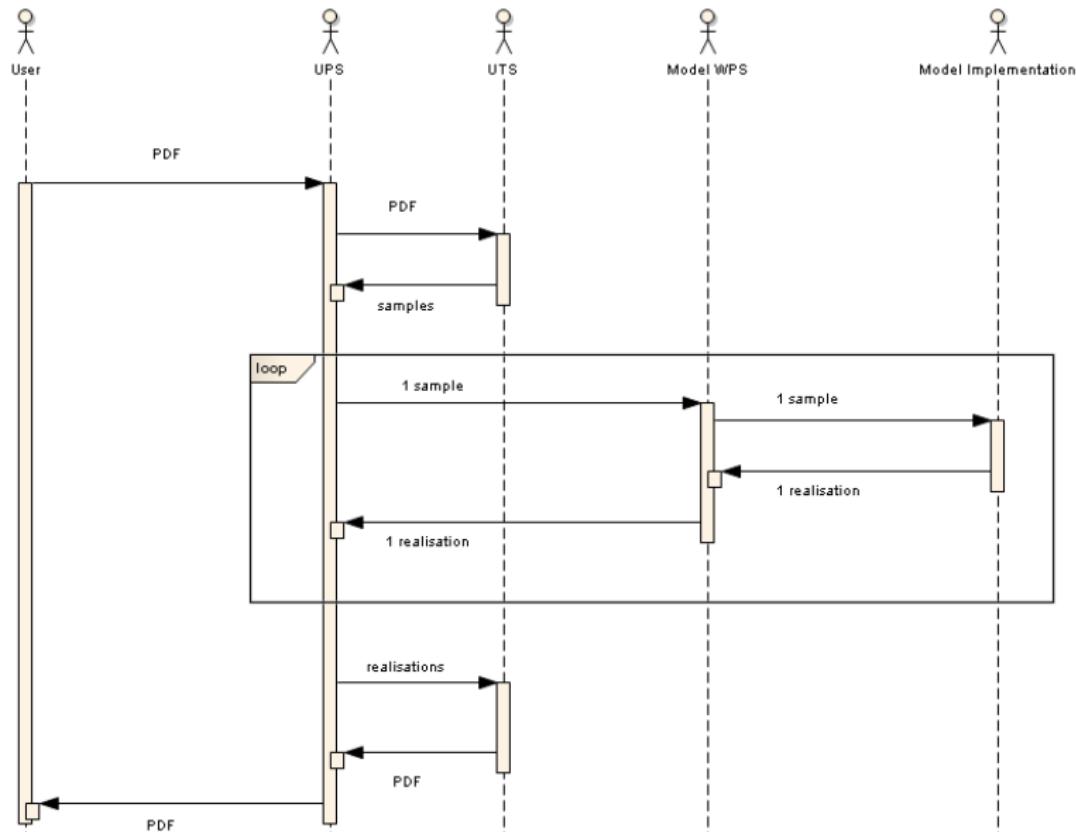
Tools: spatio-temporal aggregation service (STAS)



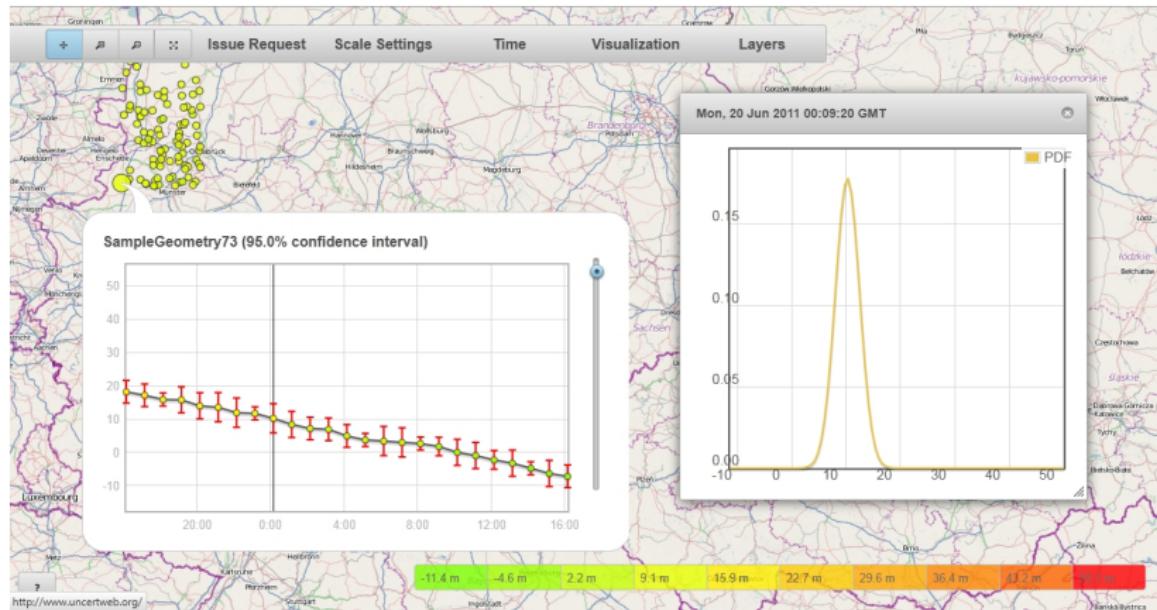
Tools: Uncertainty Proxy Service (UPS)



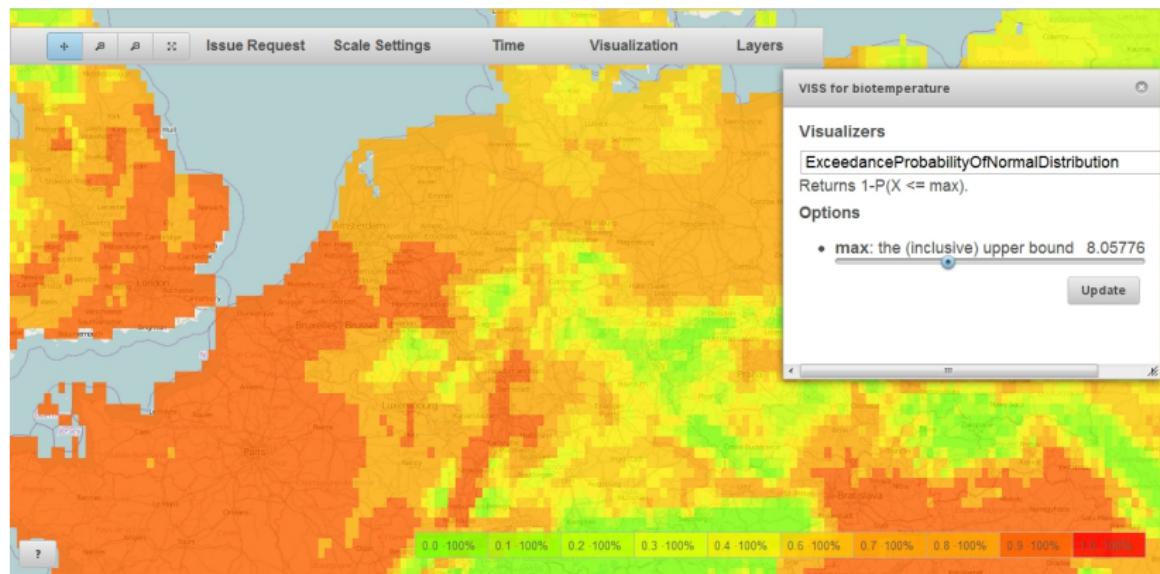
Tools: Uncertainty Transformation Service (UTS)



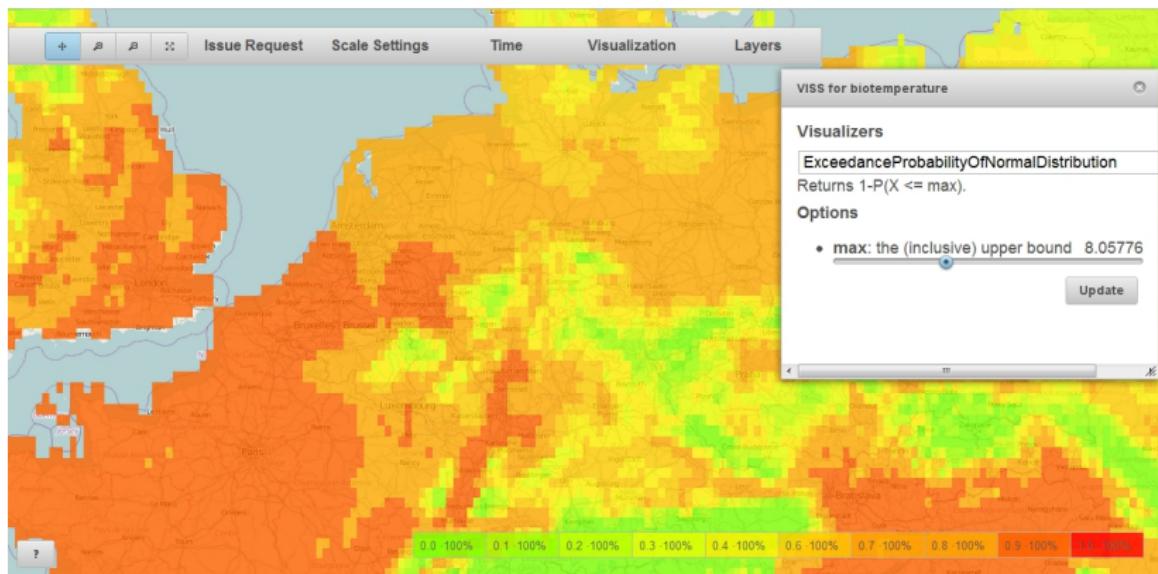
Tools: Visualizer



Tools: Visualizer



Tools: Visualizer



(usability study helped prioritizing implementation of methods)

WPS4R, R

In the background, we have WPS4R, which

- lets you upload an annotated R script
- exposes itself as a web processing service
- allows for re-use of all R functions
- (has no fancy graphics!)

But we already have ...

Modelling frameworks:

Kepler, Taverna, Trident, Vis Trails, SME, ESMF, MCT, Delta Shell, OpenMI, FRAMES, Fluid Earth, TIME, Tarsier...

Uncertainty frameworks:

PEST (Parameter ESTimation Toolkit), UCODE, OSTRICH, UNCSAM, Crystal Ball, @RISK, GENIE-1, SME (Spatial Modelling Environment), FRAMES Sensitivity/ Uncertainty module, TIME, SoftIAM, WADES, UNCSIM, DUE (Data Uncertainty Engine)...

But we already have ...

Modelling frameworks:

Kepler, Taverna, Trident, Vis Trails, SME, ESMF, MCT, Delta Shell, OpenMI, FRAMES, Fluid Earth, TIME, Tarsier...

Uncertainty frameworks:

PEST (Parameter ESTimation Toolkit), UCODE, OSTRICH, UNCSAM, Crystal Ball, @RISK, GENIE-1, SME (Spatial Modelling Environment), FRAMES Sensitivity/ Uncertainty module, TIME, SoftIAM, WADES, UNCSIM, DUE (Data Uncertainty Engine)...

yes, and we need to integrate them, and seek convergence.

The future, I: **bottom-up**, holistic

Idea:

- ① extend computing primitives (double, integer) with random variables
- ② rewrite existing code using the new primitives

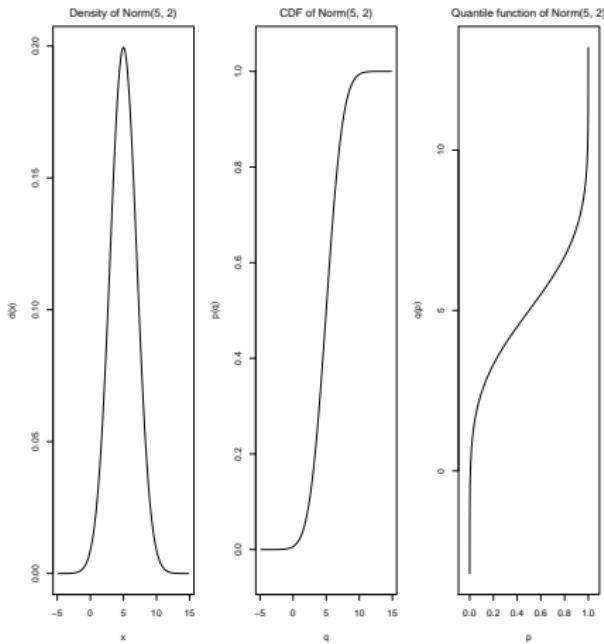
R package (family) `distr`¹ provides an *object-oriented implementation of distributions*. From `?distr`:

distr provides a conceptual treatment of distributions by means of S4 classes. A mother class Distribution is introduced with slots for a parameter and -most important- for the four constitutive methods `r`, `d`, `p`, and `q` for simulation [`r`] respectively for evaluation of density [`d`], cumulative distribution [`p`], and quantile function [`q`] of the corresponding distribution.

¹Ruckdeschel, P., Kohl, M., Stabla, T., and Camphausen, F. (2006) S4 Classes for Distributions. *R News*, 6 (2), pp 2–6.

distr

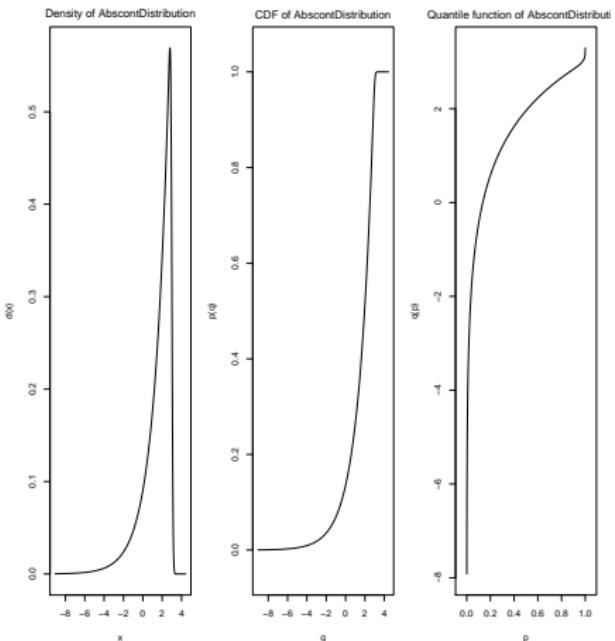
```
> library(distr)
> X = Norm(mean = 5, sd = 2)
> plot(X)
> d(X)(3.5)    # density
[1] 0.1505687
> p(X)(3.5)    # cumul. distr.
[1] 0.2266274
> q(X)(0.975)  # quantile
[1] 8.919928
> r(X)(10)      # simulate
[1] 2.6402780  0.5049697
[3] 3.6988388  4.6612276
[5] 7.1345645  3.9098073
[7] 6.8112369  10.4476685
[9] 5.8216843  4.2276650
```



distr – allows complicated expressions!

```
> X = Norm(3, 0.1)
> Y = Lnorm(3, 0.5)
> Z = Exp(2)
> P = X + sin(exp(-Y)) - 3 * Z
> plot(P)
> r(P)(5)
```

```
[1] 2.802517 2.542141 2.150656
[4] 2.822520 2.329171
```



distr – allows discrete distributions!

```
> DD <- DiscreteDistribution  
> D2 <- DD(supp = c(1:5), prob = c(0.  
+      0.2, 0.3, 0.2, 0.2))  
> d(D2)(1:5)
```

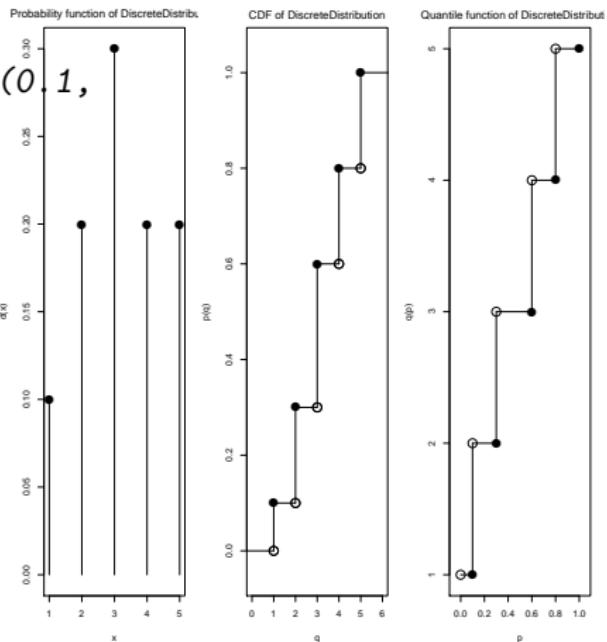
```
[1] 0.1 0.2 0.3 0.2 0.2
```

```
> p(D2)(1:5)
```

```
[1] 0.1 0.3 0.6 0.8 1.0
```

```
> plot(D2)  
> r(P)(5)
```

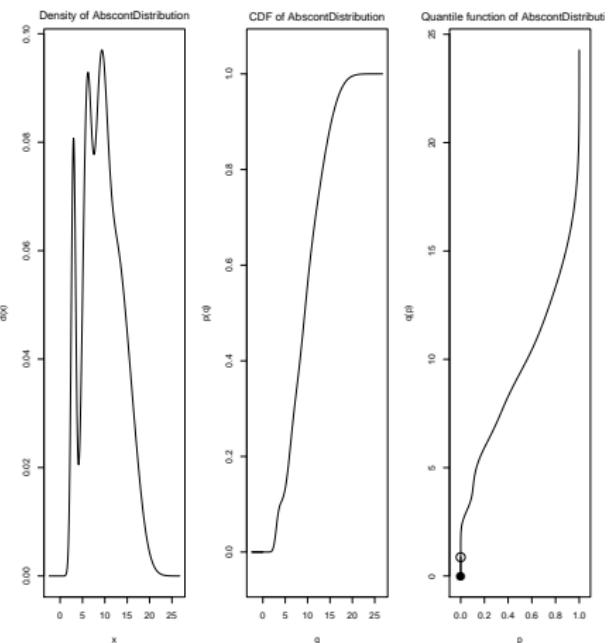
```
[1] 0.7623937 2.4667492 2.4929994  
[4] 1.3507760 0.9622070
```



distr – allows mixing discrete and continuous!

```
> Z = D2 * Norm(3, 0.5)  
> plot(Z)  
> r(Z)(10)
```

```
[1] 13.452018 2.764986 11.608150  
[4] 11.088187 3.647741 14.395616  
[7] 6.291692 13.770640 18.773374  
[10] 6.578087
```



Can distr solve the model web?

- should we recompile the models?
- how do we parameterize all the RVs?
- how to deal with multivariate distributions? (e.g. correlations in space, time or space-time)

The future, II: top-down, holistic

- model web = sensor web + models + tools
- the future: holistic computing (MC samples, S/T resolution, functional data; asynchronous, start-stop, updating of uncertainty sources)
- complexity = $f(n_{\text{time}}, n_{\text{pixels}}, n_{\text{attr}}, n_{\text{MC}})$, possibly
 $O(n_{\text{time}} \times n_{\text{pixels}} \times n_{\text{attr}} \times n_{\text{MC}})$

The future, II: top-down, holistic

- model web = sensor web + models + tools
- the future: holistic computing (MC samples, S/T resolution, functional data; asynchronous, start-stop, updating of uncertainty sources)
- complexity = $f(n_{\text{time}}, n_{\text{pixels}}, n_{\text{attr}}, n_{\text{MC}})$, possibly
$$O(n_{\text{time}} \times n_{\text{pixels}} \times n_{\text{attr}} \times n_{\text{MC}})$$
- optimize computations over [S/T resolution, attribute space, model complexity, probability space], by

The future, II: top-down, holistic

- model web = sensor web + models + tools
- the future: holistic computing (MC samples, S/T resolution, functional data; asynchronous, start-stop, updating of uncertainty sources)
- complexity = $f(n_{\text{time}}, n_{\text{pixels}}, n_{\text{attr}}, n_{\text{MC}})$, possibly

$$O(n_{\text{time}} \times n_{\text{pixels}} \times n_{\text{attr}} \times n_{\text{MC}})$$

- optimize computations over [S/T resolution, attribute space, model complexity, probability space], by
- selection/aggregation,

The future, II: top-down, holistic

- model web = sensor web + models + tools
- the future: holistic computing (MC samples, S/T resolution, functional data; asynchronous, start-stop, updating of uncertainty sources)
- complexity = $f(n_{\text{time}}, n_{\text{pixels}}, n_{\text{attr}}, n_{\text{MC}})$, possibly

$$O(n_{\text{time}} \times n_{\text{pixels}} \times n_{\text{attr}} \times n_{\text{MC}})$$

- optimize computations over [S/T resolution, attribute space, model complexity, probability space], by
- selection/aggregation, redundancy analysis,

The future, II: top-down, holistic

- model web = sensor web + models + tools
- the future: holistic computing (MC samples, S/T resolution, functional data; asynchronous, start-stop, updating of uncertainty sources)
- complexity = $f(n_{\text{time}}, n_{\text{pixels}}, n_{\text{attr}}, n_{\text{MC}})$, possibly

$$O(n_{\text{time}} \times n_{\text{pixels}} \times n_{\text{attr}} \times n_{\text{MC}})$$

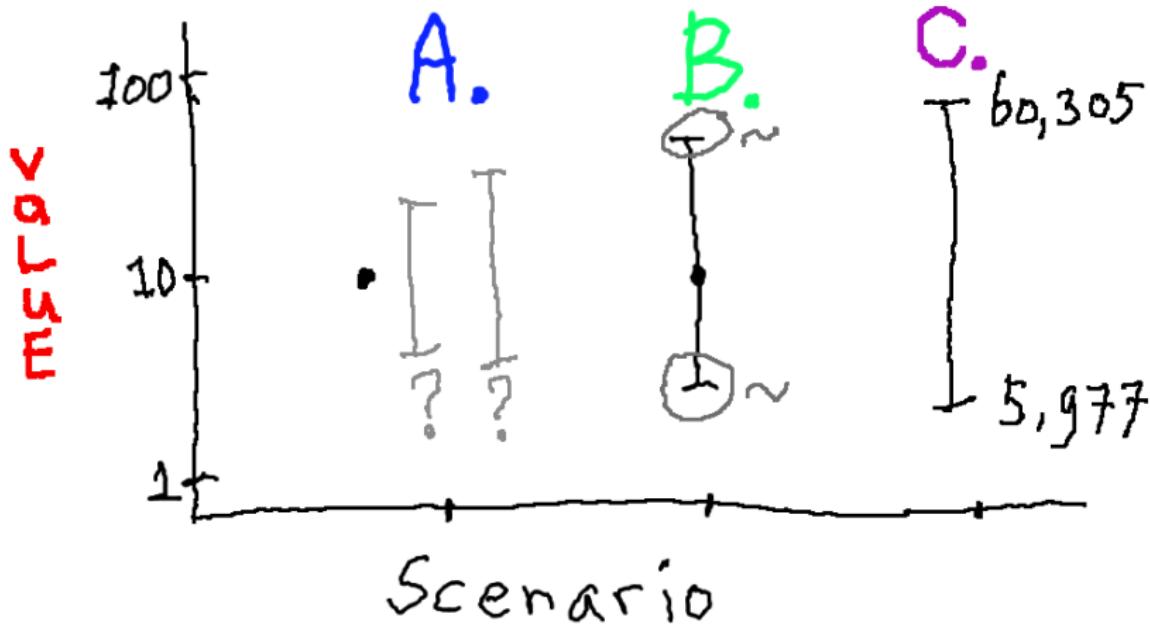
- optimize computations over [S/T resolution, attribute space, model complexity, probability space], by
- selection/aggregation, redundancy analysis, emulation,

The future, II: top-down, holistic

- model web = sensor web + models + tools
- the future: holistic computing (MC samples, S/T resolution, functional data; asynchronous, start-stop, updating of uncertainty sources)
- complexity = $f(n_{\text{time}}, n_{\text{pixels}}, n_{\text{attr}}, n_{\text{MC}})$, possibly

$$O(n_{\text{time}} \times n_{\text{pixels}} \times n_{\text{attr}} \times n_{\text{MC}})$$

- optimize computations over [S/T resolution, attribute space, model complexity, probability space], by
- selection/aggregation, redundancy analysis, emulation, smart MC sampling



Wrapping up – conclusions

- we ought to use R,

Wrapping up – conclusions

- we ought to use R,
- the model web is there to come, and we'll uncertainty-enable it before it arrives,

Wrapping up – conclusions

- we ought to use R,
- the model web is there to come, and we'll uncertainty-enable it before it arrives,
- my future visions are holistic, probably bottom-up *and* top-down,

Wrapping up – conclusions

- we ought to use R,
- the model web is there to come, and we'll uncertainty-enable it before it arrives,
- my future visions are holistic, probably bottom-up *and* top-down,
- spatial data quality is not only booming, it's also fun!

Thanks!

Credits go to:

UncertWeb team: Lucy Bastin, Dan Cornford, Gregoire Dubois, Gerard Heuvelink, Richard Jones, Paolo Mazzetti, Stefano Nativi, Jon Skøien, Matt Williams, ...

ifgi: Merret Buurman, Lydia Gerharz, Kristina Helle, Matthias Hinz, Daniel Nüst, Benjamin Proß, Hansi Senaratne, Christoph Stasch, ...