From Open Source to Open Science

Markus Neteler
Fondazione E. Mach – CRI, Italy
http://gis.cri.fmach.it

In collaboration with:
Markus Metz, Duccio Rocchini, Luca Delucchi, FEM
Luigi Ponti, ENEA http://utagri.enea.it - CASAS http://cnr.berkeley.edu/casas

FOSS4G-CEE Conference

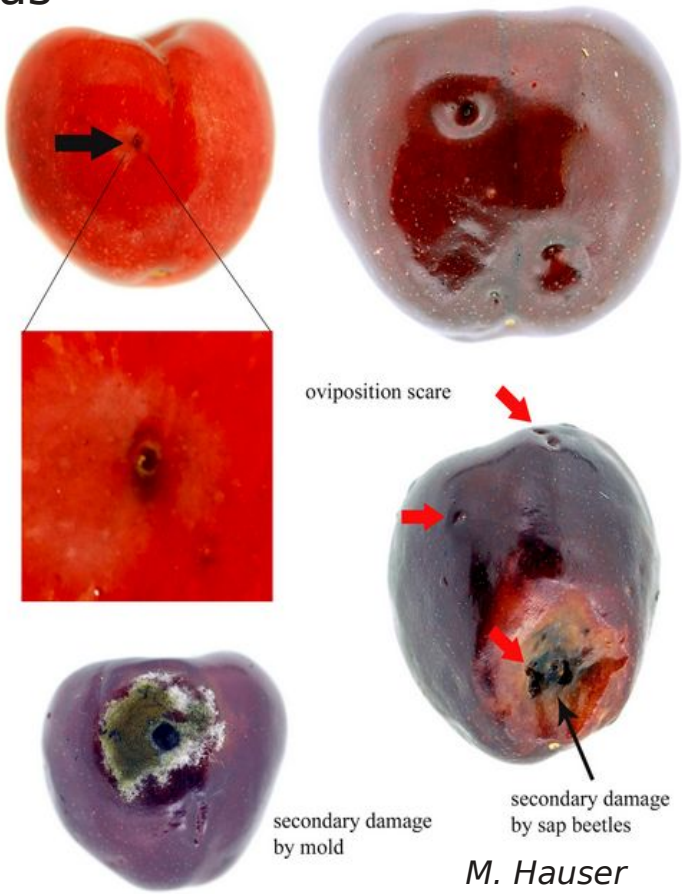
21-23 May 2012, Prague, Czech Republic

Diseases and parasites:
New global challenges arriving

Drosophila suzukii (Spotted-wing drosophila - SWD, vinegar fly)

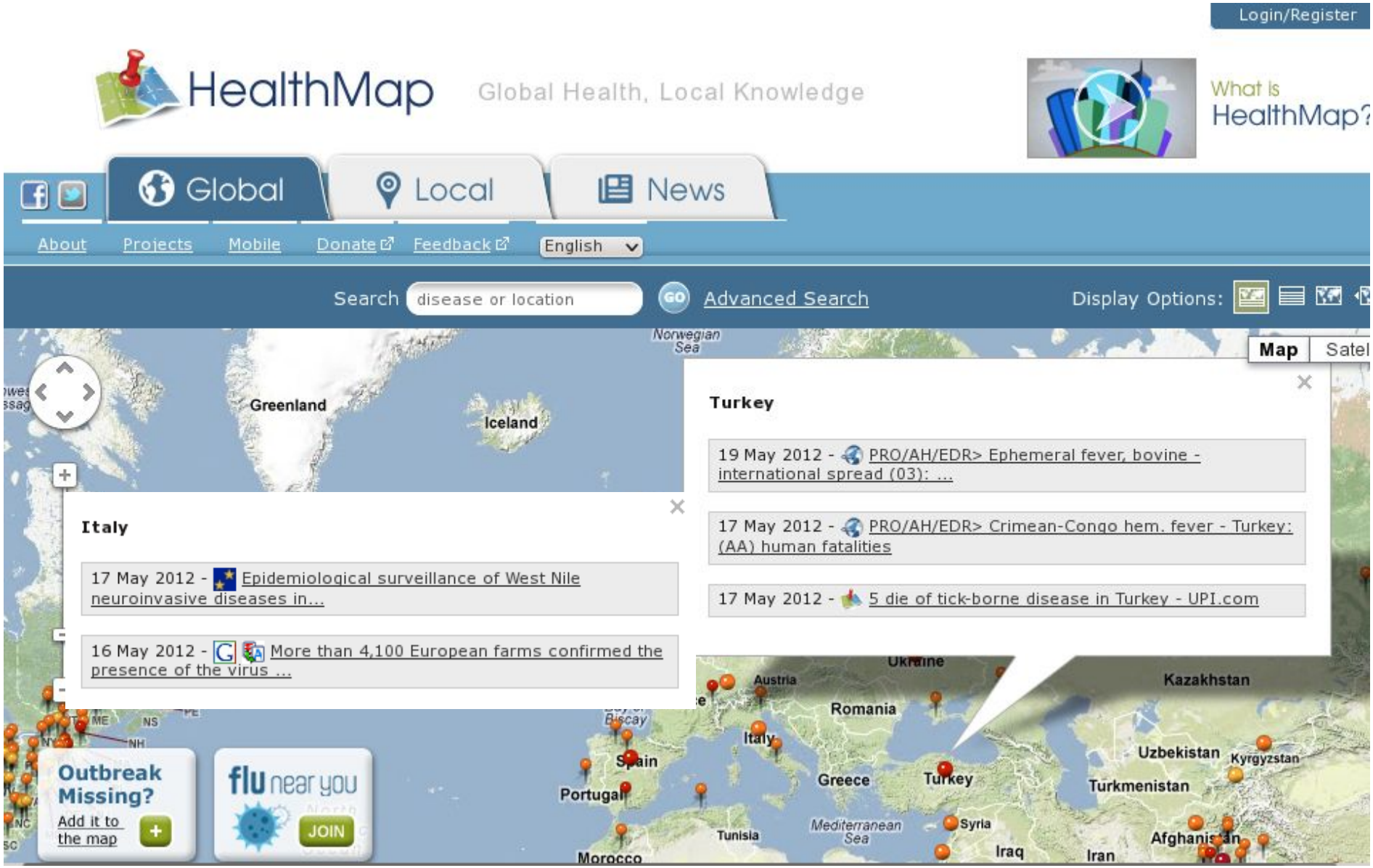
- new invasive insect species
- threatening **US and European fruit production**
- infests unwounded **ripening fruits:**
  berries, stone fruits, grapes and damaged fruits as loquat, persimmons, and tomatoes.
- extreme fecundity and short generation times
- huge fruit losses already in million Euro range

Photo: M. Hauser



M. Hauser

Emerging zoonotic diseases in Europe

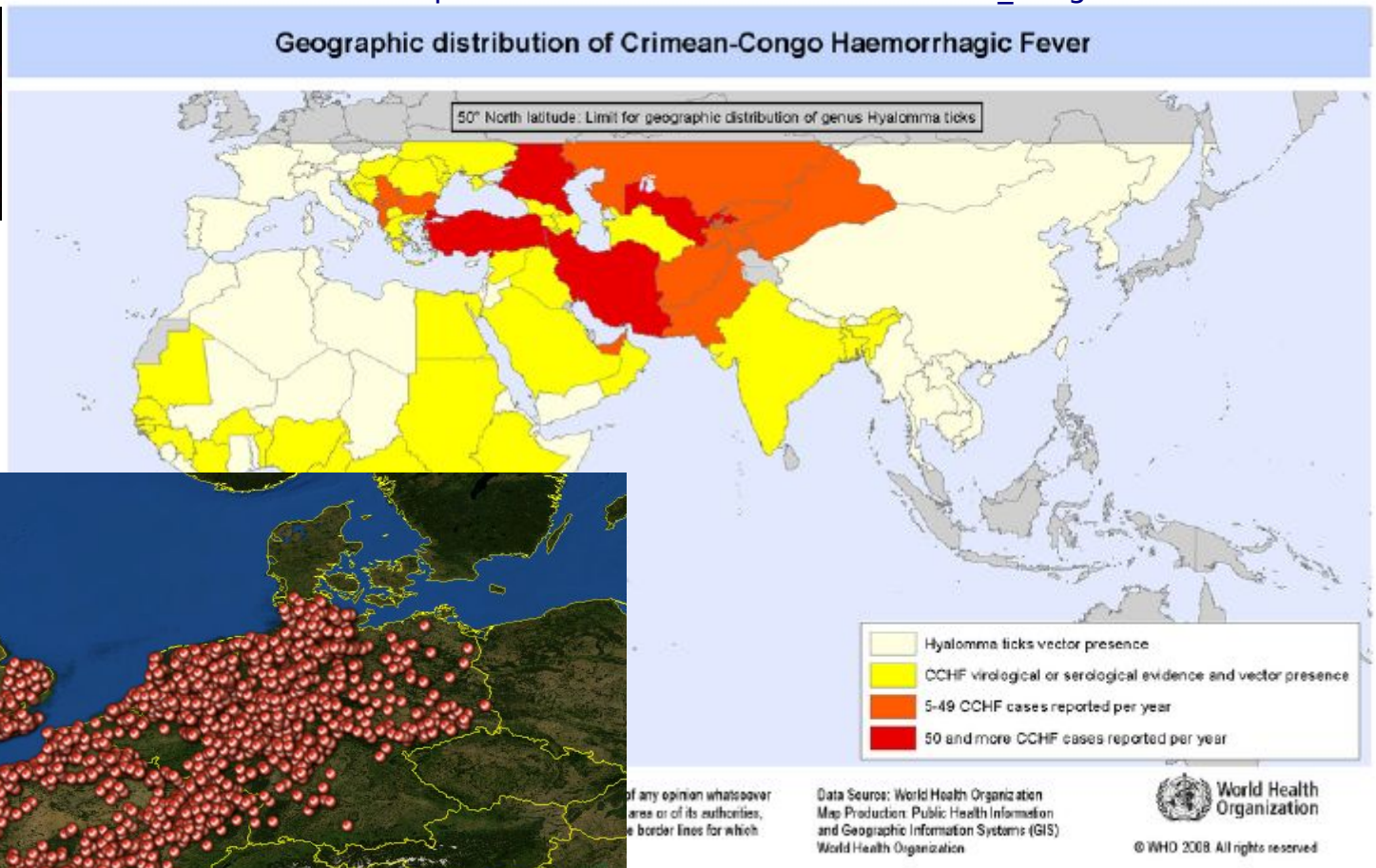


Map of 19 May 2012

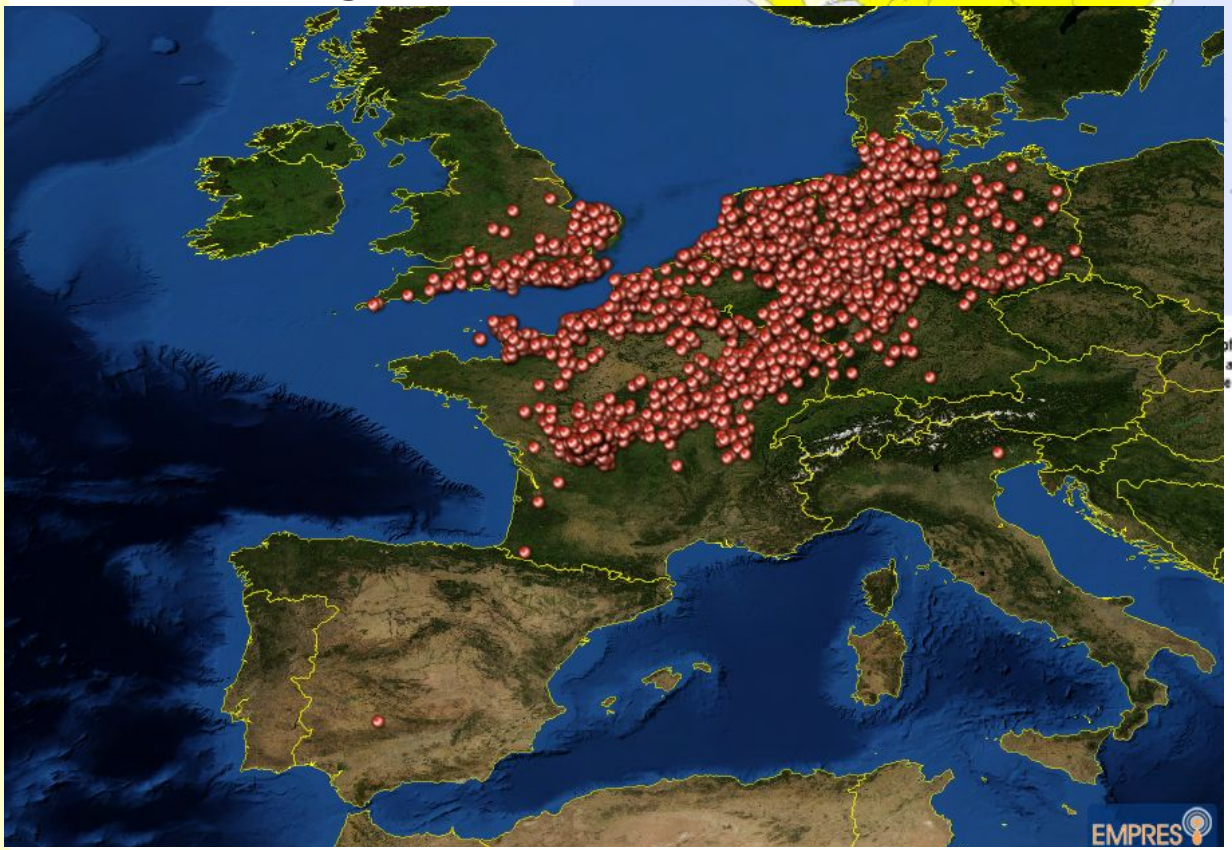
http://healthmap.org/en/

Emerging zoonotic diseases in Europe

http://www.who.int/csr/disease/crimean_congoHF/en/index.html



Schmallenberg virus



http://www.flutrackers.com/forum/showthread.php?t=183181

Ecology… an open(ing) science

Review

Cell PRESS

Trends in Ecology and Evolution
February 2012, Vol. 27, No. 2

Special Issue: Ecological and evolutionary informatics

Ecoinformatics: supporting ecology as a data-intensive science

William K. Michener[1] and Matthew B. Jones[2]

[1] University Libraries, University of New Mexico, Albuquerque, NM 87131, USA
[2] National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, Santa Barbara, CA 93101, USA

Ecology is evolving rapidly and increasingly changing into a more open, accountable, interdisciplinary, collaborative and data-intensive science. Discovering, integrating and analyzing massive amounts of heterogeneous data are central to ecology as researchers address complex questions at scales from the gene to the biosphere. Ecoinformatics offers tools and approaches for managing ecological data and transforming the data into information and knowledge. Here, we review the state-of-the-art and recent advances in ecoinformatics that can benefit ecologists and environmental scientists as they tackle increasingly challenging questions that require voluminous amounts of data across disciplines and scales of space and time. We also highlight the challenges and opportunities that remain.

Data be free!

run on powerful distributed computing systems. For example, Kepler includes facilities for easily executing models on pre-existing computing grids, in cloud-computing environments and in ad hoc networks of workflow systems [65,66], while capturing a full provenance trace of the process; and VisTrails is built to generate effectively scientific visualizations while also capturing the provenance of the analysis [61].

Supporting the full data life cycle

New ground, aerial and satellite-based environmental observing systems coupled with the rapid growth in the use of in situ environmental sensor networks for field research and monitoring, as well as an ever-growing number of citizen-science programs, will soon push ecology and the environmental sciences into a new era where petabytes of data are being collected annually. Powerful informatics platforms will be required to support scientists as they move into this age of data-intensive science. Several such platforms are being designed and built at various scales, including the LTER NIS, the DataONE Federation, LifeWatch, NEON, GLEON and OOI.

The US LTER Network is presently building a network information system that will support synthetic science by: (i) using standardized metadata management and access approaches; (ii) providing middleware programs and workflow solutions that facilitate the creation and maintenance of integrated LTER data sets; and (iii) supporting standardized applications that facilitate discovery, access and use of LTER data [25,67].

DataONE represents a new type of research platform

Box 3. Open science for society

Global problems require open access to global data from many disciplines. Such data arise from scientific disciplines that often have very different cultures with respect to data sharing, development and adoption of standards, and practice of good data stewardship. Incentives from research sponsors, societies and institutions (e.g. requiring data management plans) combined with the availability of new informatics tools and platforms, such as DataONE, will be necessary to facilitate data intensive science. Three avenues of research and development offer particular promise: (i) automated provenance-tracking mechanisms that allow scientists to understand and replicate scientific findings fully [76]; (ii) advanced visual analytics that enable scientists to interpret complex, large data volumes more rapidly [68]; and (iii) usability analysis and software engineering support that enable scientists to use advanced ecoinformatics tools more easily.

Tracking the provenance of scientific results is particularly important as advances in environmental science are applied to issues important to society. Open data provide the feedstock on which good science is based, replicable analysis and modeling practices lead to robust findings, and open-access publication disseminates these critical results to the broadest audiences, ensuring the greatest impact of open science for society.

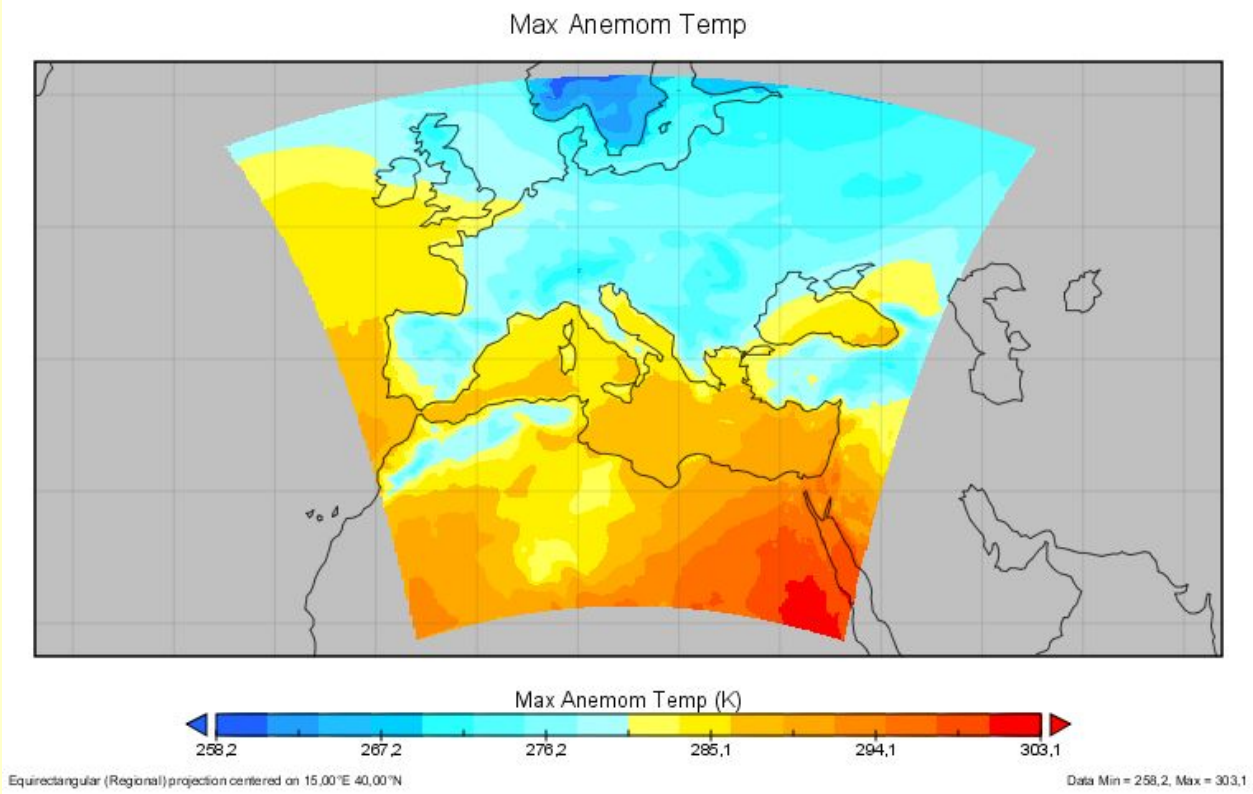
research must be openly available and the approaches used in deriving scientific findings must be transparent to ensure that science and society maximally benefit (Box 3).

Remaining challenges

Despite the emergence of ecoinformatics solutions that enable science, several technical and sociocultural challenges and research opportunities remain. First, from the technical side, it is difficult to transport terabyte- and petabyte-sized data sets. Possible solutions include adding

Analysis of Mediterranean olive systems using the PROTHEUS present climate data

By Luigi Ponti, ENEA, Italy



ERA-40 reanalysis climate data for 1958-2000

↓

PROTHEUS: Regional climate model RegCM3 coupled to MIT ocean model

↓

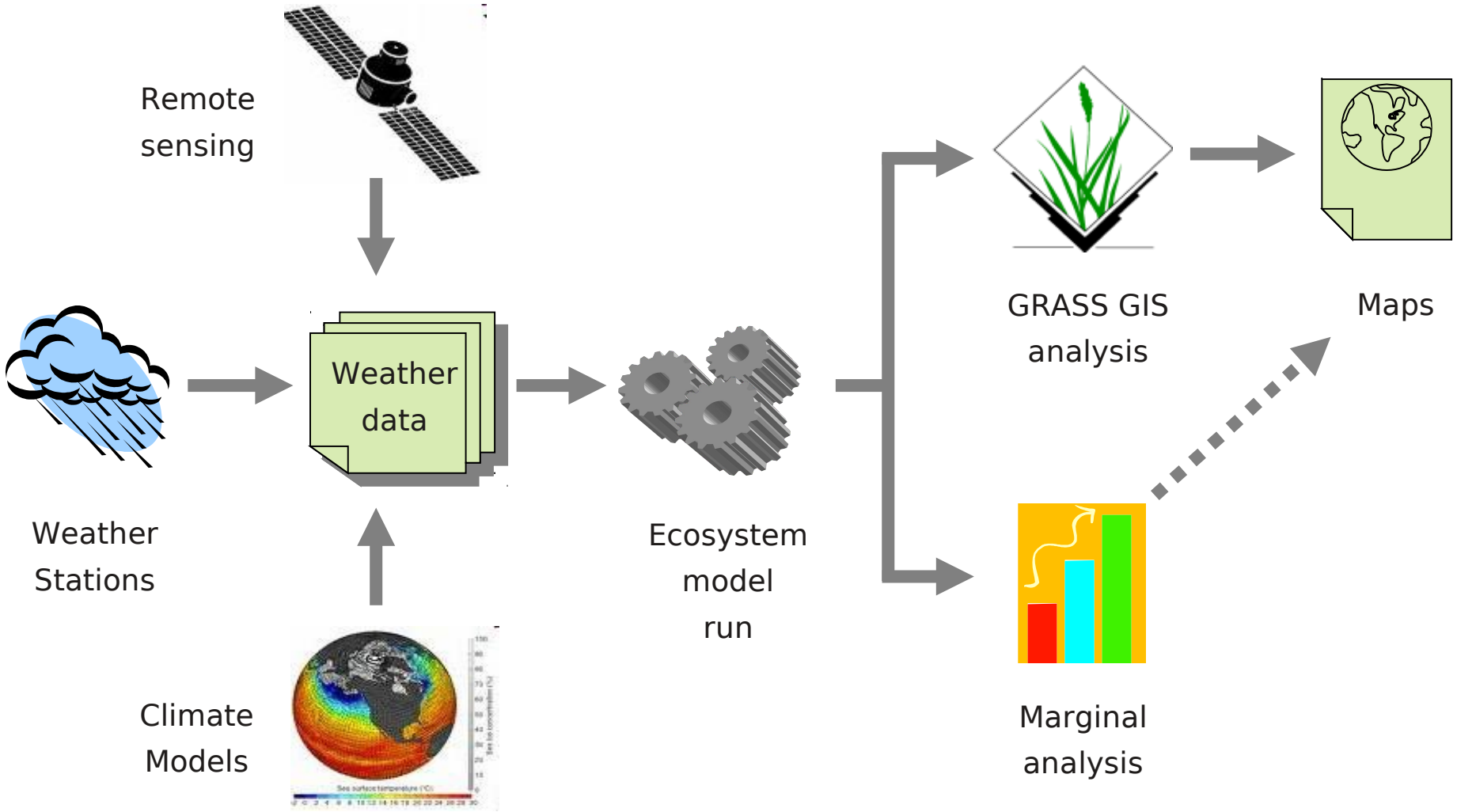
Down-scaling of climate data for the Mediterranean region

The GlobalChangeBiology project
Climate change & Mediterranean agroecosystems
http://sites.google.com/site/globalchangebiology/

Artale et al. 2010

Ecosystem analysis: integration of site-specific weather data, GIS maps and marginal analysis

By Luigi Ponti, ENEA, Italy

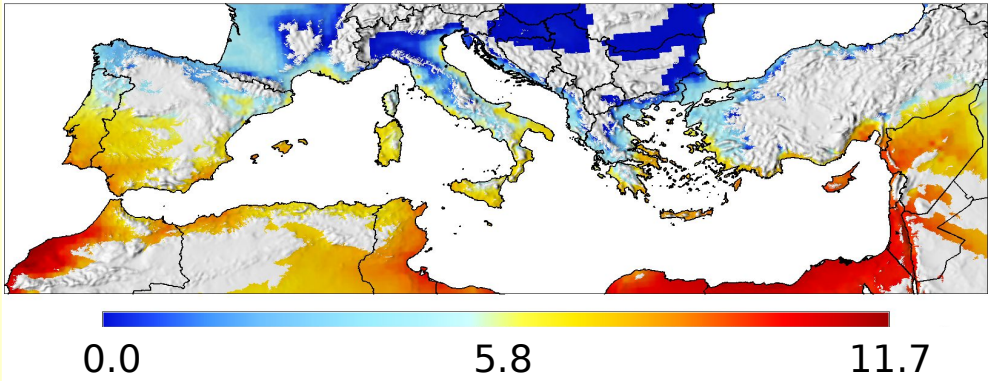


Gutierrez et al. 2010

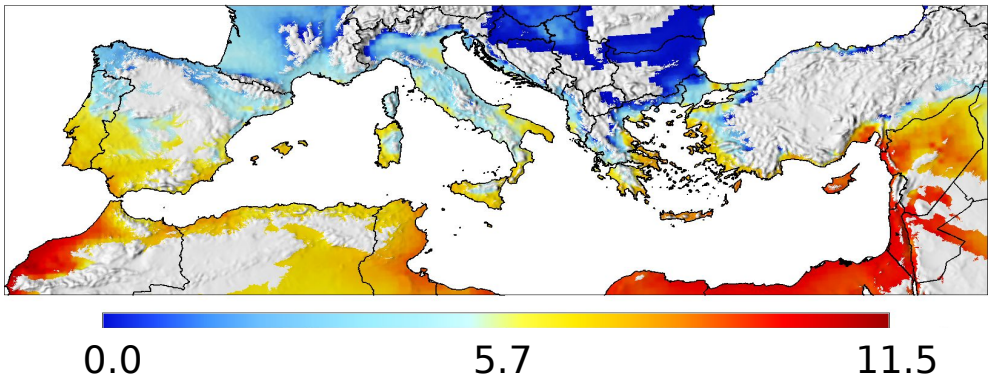
Multitrophic interactions of olive and olive fly mapped across the Mediterranean

By Luigi Ponti, ENEA, Italy

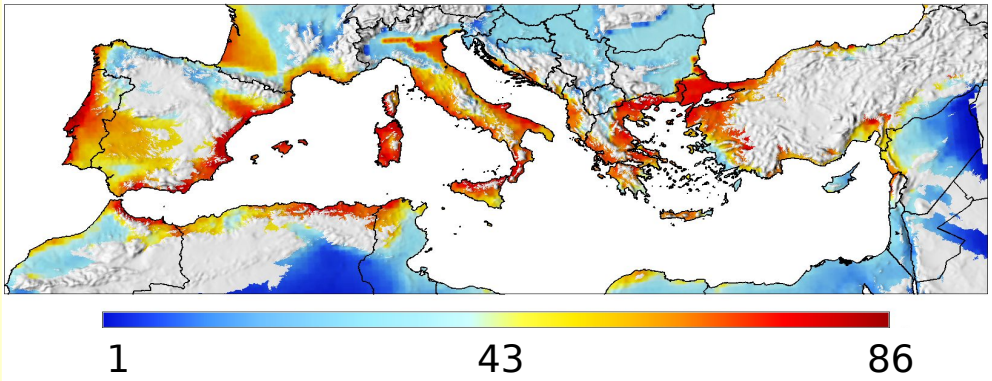
Average olive yield (kg), 1958-1967



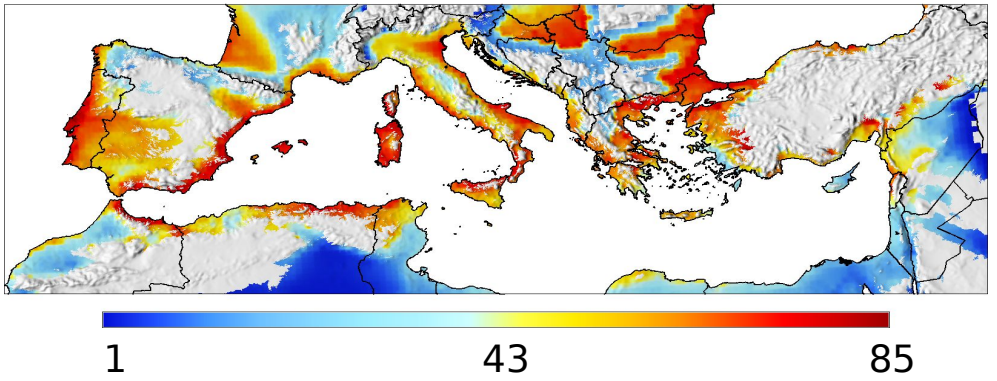
Average olive yield (kg), 1988-1997



% fruit attacked by olive fly, 1958-1967



% fruit attacked by olive fly, 1988-1997

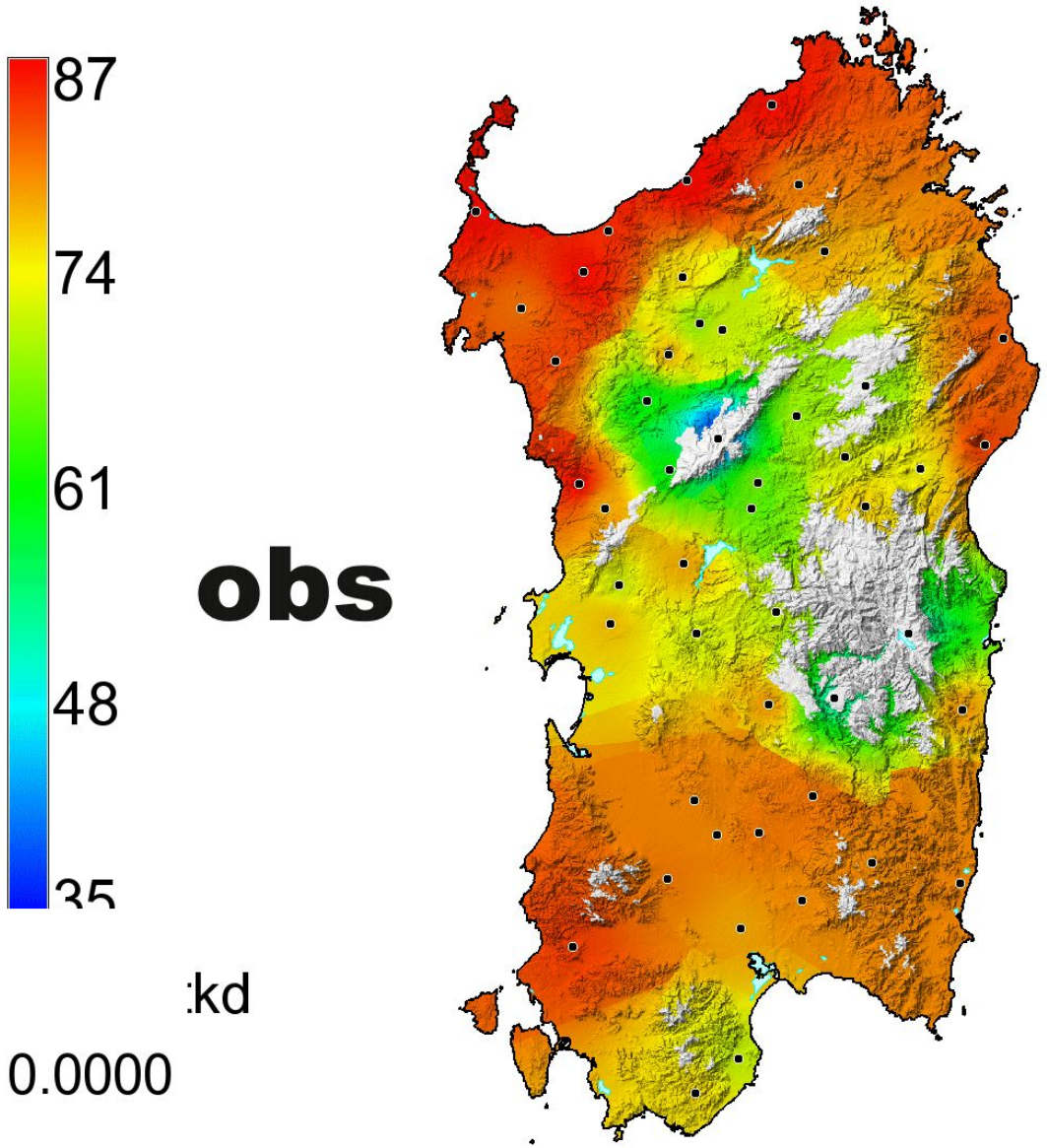


Ponti et al. 2009

The GlobalChangeBiology project
Climate change & Mediterranean agroecosystems
http://sites.google.com/site/globalchangebiology/

Olive fly infestation % in Sardinia under climate warming

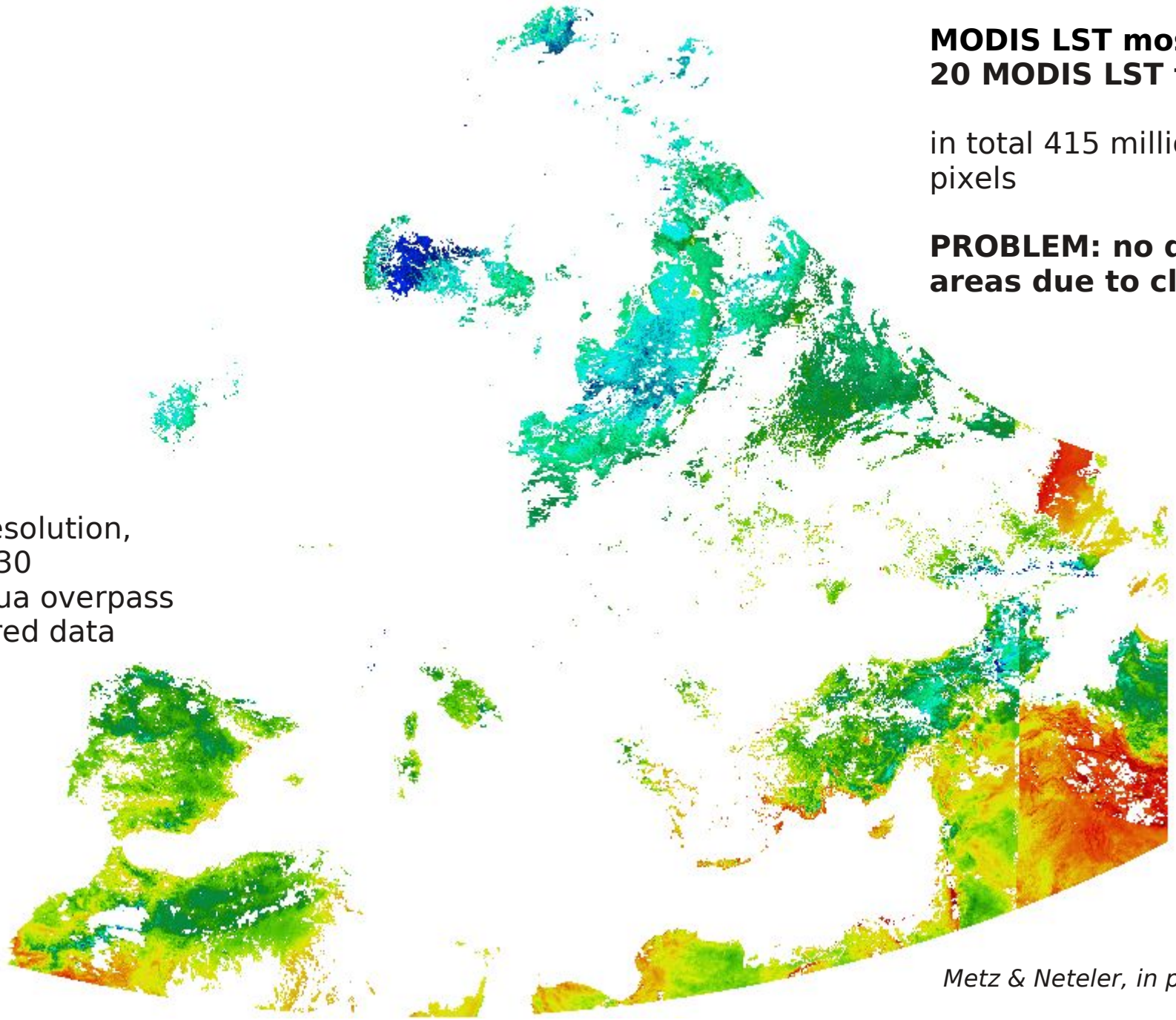
By Luigi Ponti, ENEA, Italy



The GlobalChangeBiology project
Climate change & Mediterranean agroecosystems
http://sites.google.com/site/globalchangebiology/

Ponti et al. 2009

MODIS LST at European scale (filtered mosaic)



MODIS LST mosaic of 20 MODIS LST tiles

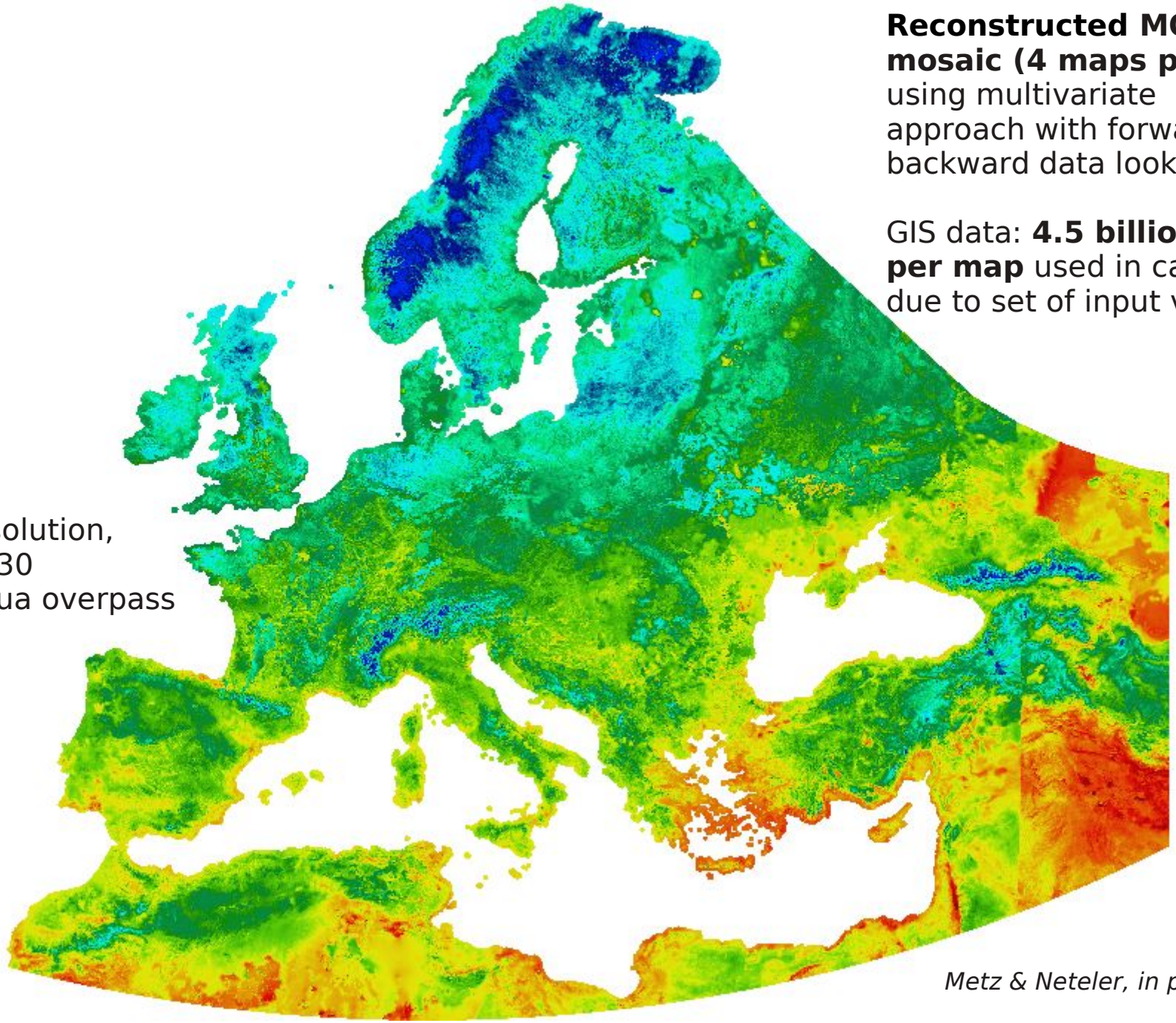
in total 415 million pixels

PROBLEM: no data areas due to clouds

1000m resolution,
2010-05-30
01:30 Aqua overpass
Raw-filtered data

Metz & Neteler, in prep.

MODIS LST at European scale (reconstructed)



Reconstructed MODIS LST mosaic (4 maps per day) using multivariate approach with forward/ backward data lookup

GIS data: **4.5 billion pixels per map** used in calculation due to set of input variables

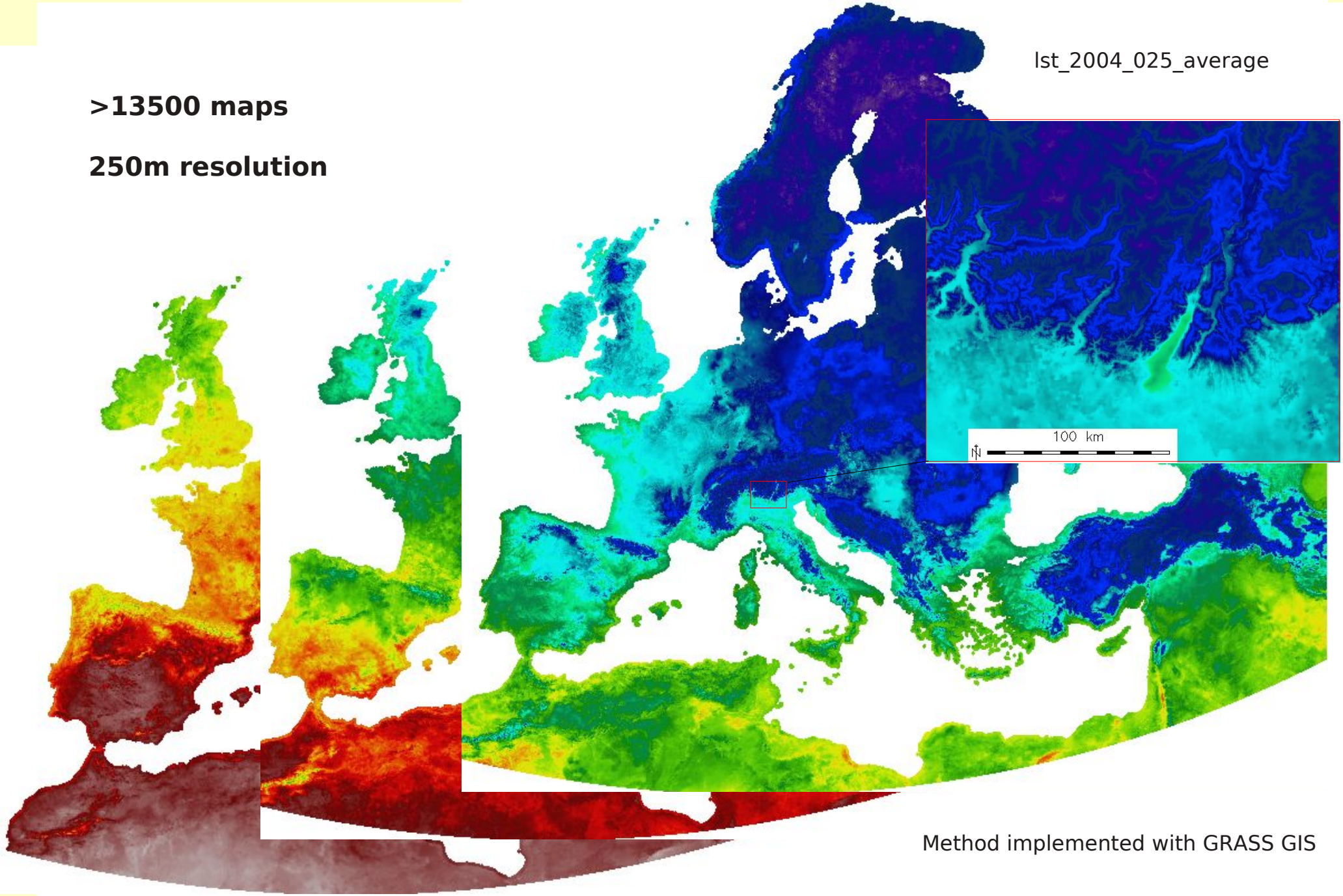
250m resolution,
2010-05-30
01:30 Aqua overpass
gap-filled

Metz & Neteler, in prep.

MODIS LST at European scale (reconstructed)

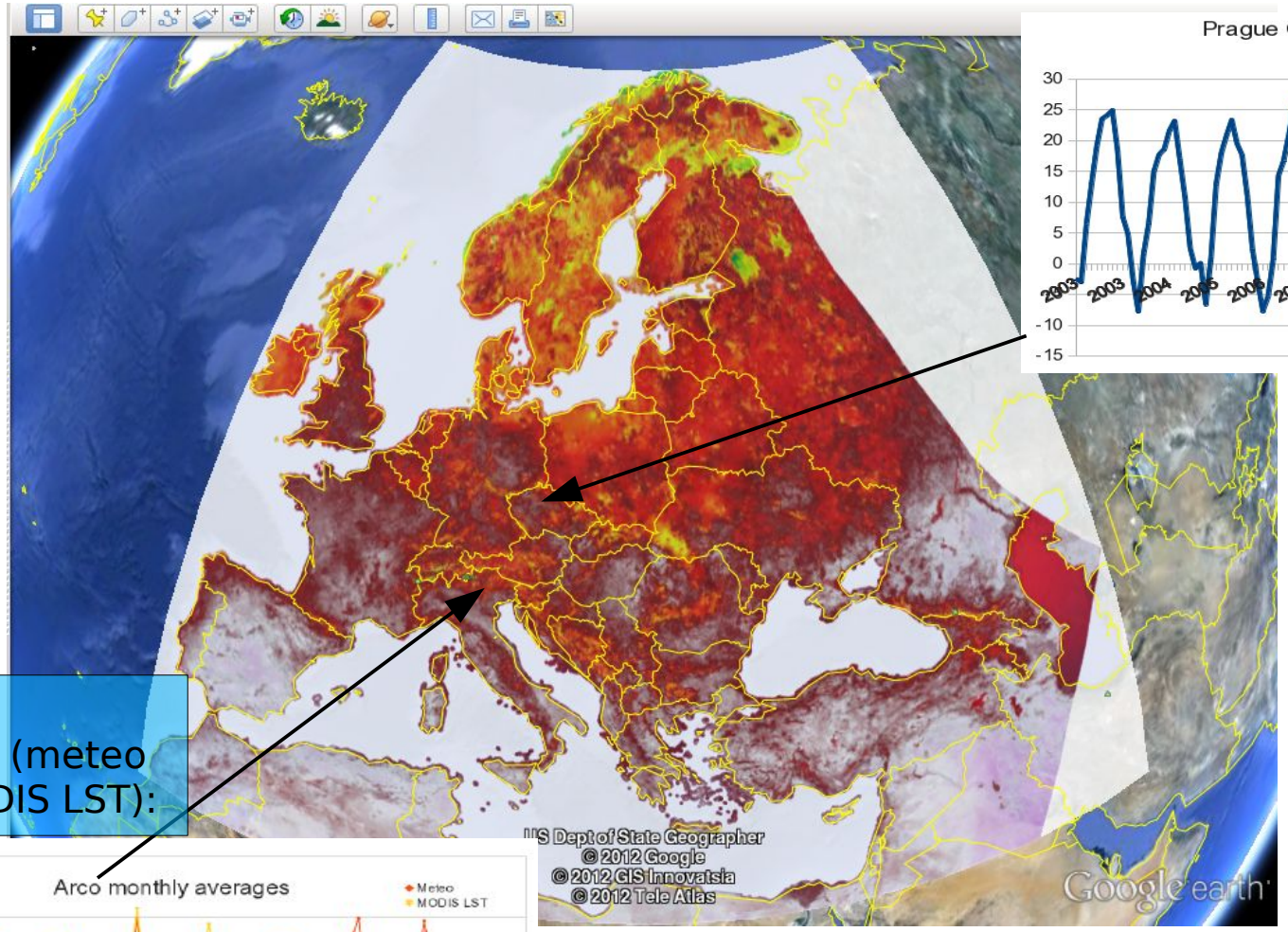
>13500 maps

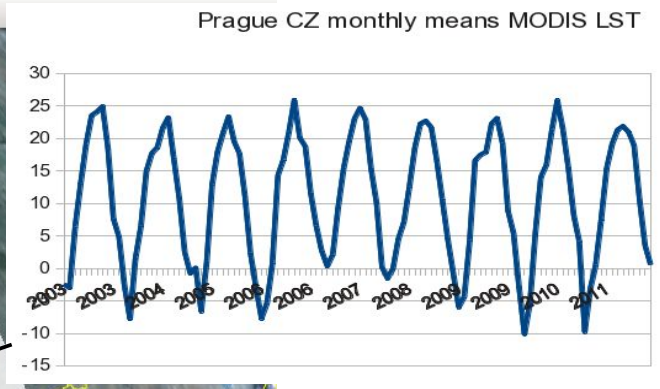
250m resolution



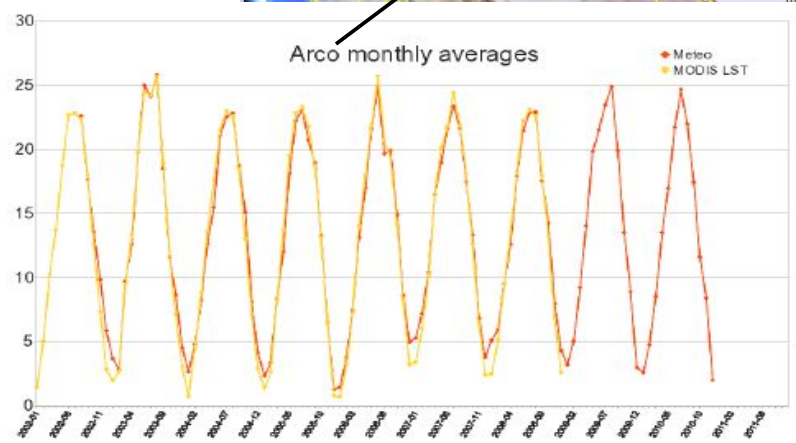
Method implemented with GRASS GIS

The new European daily MODIS LST time series





Pixel-wise time series (meteo versus MODIS LST):



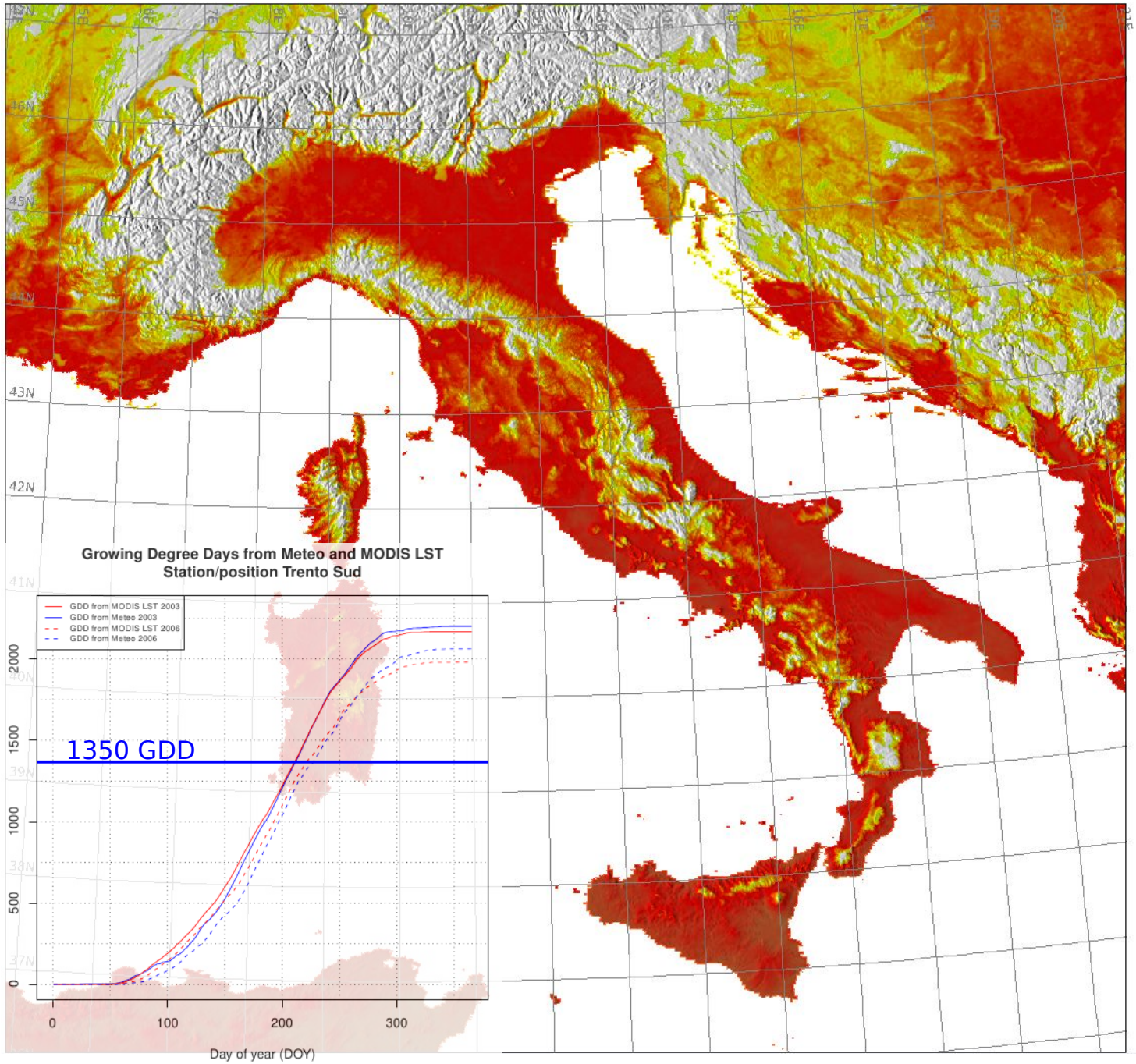
European MODIS LST mosaic

... usable as **virtual meteorologic stations for temperature**

250m resolution
4 maps per **day**
data since 2000

Metz & Neteler, in prep.

Threshold map >1350 GDD from MODIS LST



GDD.2010_b11

3135

2351

1567

Threshold of 1350 GDD after
Kobayashi et al., 2002.
J Med Entomol, 39:4-11.

Neteler, Metz, in prep.

Implemented in GRASS GIS

Ecology… an open(ing) science

Review

Cell PRESS

Trends in Ecology and Evolution
February 2012, Vol. 27, No. 2

Special Issue: Ecological and evolutionary informatics

Ecoinformatics: supporting ecology as a data-intensive science

William K. Michener[1] and Matthew B. Jones[2]

[1] University Libraries, University of New Mexico, Albuquerque, NM 87131, USA
[2] National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, Santa Barbara, CA 93101, USA

Ecology is evolving rapidly and increasingly changing into a more open, accountable, interdisciplinary, collaborative and data-intensive science. Discovering, integrating and analyzing massive amounts of heterogeneous data are central to ecology as researchers address complex questions at scales from the gene to the biosphere. Ecoinformatics offers tools and approaches for managing ecological data and transforming the data into information and knowledge. Here, we review the state-of-the-art and recent advances in ecoinformatics that can benefit ecologists and environmental scientists as they tackle increasingly challenging questions that require voluminous amounts of data across disciplines and scales of space and time. We also highlight the challenges and opportunities that remain.

Data be free!

Yes, but what's still missing…?

run on powerful distributed computing systems. For example, Kepler includes facilities for easily executing models on pre-existing computing grids, in cloud-computing environments and in ad hoc networks of workflow systems [65,66], while capturing a full provenance trace of the process; and VisTrails is built to generate effectively sci-

Supporti

New ground, aerial and satellite-based environmental observing systems coupled with the rapid growth in the use of in situ environmental sensor networks for field research and monitoring, as well as an ever-growing number of citizen-science programs, will soon push ecology and the environmental sciences into a new era where petabytes of data are being collected annually. Powerful informatics platforms will be required to support scientists as they move into this age of data-intensive science. Several such platforms are being designed and built at various scales, including the LTER NIS, the DataONE Federation, LifeWatch, NEON, GLEON and OOI.

The US LTER Network is presently building a network information system that will support synthetic science by: (i) using standardized metadata management and access approaches; (ii) providing middleware programs and workflow solutions that facilitate the creation and maintenance of integrated LTER data sets; and (iii) supporting standardized applications that facilitate discovery, access and use of LTER data [25,67].

DataONE represents a new type of research platform

Box 3. Open science for society

Global problems require open access to global data from many disciplines. Such data arise from scientific disciplines that often have very different cultures with respect to data sharing, development and adoption of standards, and practice of good data stewardship. Incentives from research sponsors, societies and institutions (e.g. requiring data management plans) combined with the availability of new informatics tools and platforms, such as DataONE, will be necessary to facilitate data intensive science. Three avenues of research and development offer particular promise: (i) automated provenance-tracking mechanisms that allow scientists to understand and replicate scientific findings fully [76]; (ii) advanced visual analytics that enable scientists to interpret complex, large data volumes more rapidly [68]; and (iii) usability analysis and software engineering support that enable scientists to use advanced ecoinformatics tools more easily.

Tracking the provenance of scientific results is particularly important as advances in environmental science are applied to issues important to society. Open data provide the feedstock on which good science is based, replicable analysis and modeling practices lead to robust findings, and open-access publication disseminates these critical results to the broadest audiences, ensuring the greatest impact of open science for society.

research must be openly available and the approaches used in deriving scientific findings must be transparent to ensure that science and society maximally benefit (Box 3).

Remaining challenges

Despite the emergence of ecoinformatics solutions that enable science, several technical and sociocultural challenges and research opportunities remain. First, from the technical side, it is difficult to transport terabyte- and petabyte-sized data sets. Possible solutions include adding

Open science wants Open Source!

Trends in Ecology and Evolution June 2012, Vol. 27, No. 6

Let the four freedoms paradigm apply to ecology

Duccio Rocchini and Markus Neteler

Fondazione Edmund Mach, Research and Innovation Centre, Department of Biodiversity and Molecular Ecology, Via E. Mach 1, 38010 S. Michele all'Adige (TN), Italy

In 1985, Richard Stallman, one of the most brilliant minds in computer science, founded the Free Software Foundation and launched the concept of 'copyleft', the opposite of copyright. The aim, outlined in the GNU Manifesto (http://www.gnu.org/gnu/manifesto.html, [1]), was to make software programs 'free' as in 'freedom'.

The famous 'four freedoms' expounded by Stallman [1] are: (i) the freedom to run the program for any purpose; (ii) the freedom to study how the program works and adapt it to one's own needs; (iii) the freedom to redistribute copies; and (iv) the freedom to make improvements to the program and release them to the public. Thus, the whole (scientific) community benefits from software development. These freedoms are also inherent in several free software licenses, the GNU General Public License (GPL) being one of the most popular.

Approximately a quarter of a century after Stallman put forward his ideas, William K. Michener and Matthew B. Jones, in an article in *TREE* [2] focusing on the analysis of ecological data, stated that: 'analytical processes are fundamental to most published results in ecology'. Explicit reference to the analytical procedures adopted in generating scientific results is crucial for reproducibility, yet these processes are rarely documented in published ecological papers [2]. Scientific workflow applications, such as Kepler (https://kepler-project.org), attempt to address the problem [2], but are only partially successful because the underlying algorithms may still be opaque.

In our view, the explicit use of Free and Open Source Software (FOSS) with availability of the code is essential for completely open science: 'scientific communication relies on evidence that cannot be entirely included in publications', but 'anything less than the release of source programs is intolerable for results that depend on computation' [3].

The idea of FOSS and the public availability of the code has been around for almost as long as software [4]. Nonetheless, as far as ecologists are concerned, the open source philosophy is only just taking off, as Stokstad has also pointed out [5].

The increasing availability of open ecological data through networks such as the Global Biodiversity Information Facility (GBIF, http://www.gbif.org, [6]) or the Data Observation Network for Earth (DataONE) federated data archive (http://www.dataone.org, [7]) makes it increasingly possible to test cutting-edge ecological theories, such as dark diversity [8], evolutionary paths [9] and climate change scenarios [10]. In using a shared open-source code for testing these ecological theories, researchers can be sure that their results are reliable and also that the code they have used is robust [11]. This is particularly true when complex algorithms (or statistical approaches) are involved.

To avoid black box calculations and built-in user interfaces, criticized in [2], researchers have recourse to several examples of FOSS in areas of ecological research, such as ecological statistics (e.g. R Language and Environment for Statistical Computing, http://www.R-project.org, [12]) and spatial ecology [e.g. Geographical Resources Analysis Support System (GRASS) GIS, http://grass.osgeo.org, [4]). The modular design of such software means decentralized contributions can be made to the source code and allows different institutions and individuals around the world to improve the code base.

If FOSS were more widely employed in ecology and the code used in data analysis provided in scientific papers, more researchers [11] would be able to rely on and replicate peer-reviewed functions. Efforts still need to be made in this area to improve the processes for sharing what is in effect the backbone of ecological software: its code. Therefore, there is an urgent need to embrace Stallman's four freedoms paradigm in ecology.

Corresponding author: Rocchini, D. (duccio rocchini@gmail.com), (duccio.rocchini@fmach.it).

Acknowledgments

We would like to thank Anne Ghisla, Luca Delucchi and Tessa Say for valuable suggestions. DR is partially funded by the Autonomous Province of Trento (Italy) within the ACE-SAP project (University and Scientific Research Service regulation number 23, June 12, 2008).

Why we are here at FOSS4G-CEE...

Conclusion: Open science wants Open Source!

nature International weekly journal of science

nature news home | news archive | specials | opinion | features | news blog

comments on this story

Stories by subject

- Lab life

Published online 13 October 2010 | *Nature* **467**, 753 (2010) | doi:10.1038/467753a

Column: World View

Publish your computer code: it is good enough

Freely provided working code — whatever its quality — improves programming and enables others to engage with your research, says Nick Barnes.

Nick Barnes

I am a professional software engin
secret with scientists: most profes
very good. The code inside your la
often badly documented, inconsist

Why does this matter to science? E
published research papers often re
which means that most scientists
scientists generally think the code

"If you are still hesitant about releasing your code, then ask yourself this question: **does it perform the algorithm you describe in your paper**? If it does, your audience will accept it, and maybe feel happier with its own efforts to write programs. If not, well, you should fix that anyway."

[Nick Barnes]

Let's add to this:

YES, but aim to integrate your code into an Open Source **Community** project!

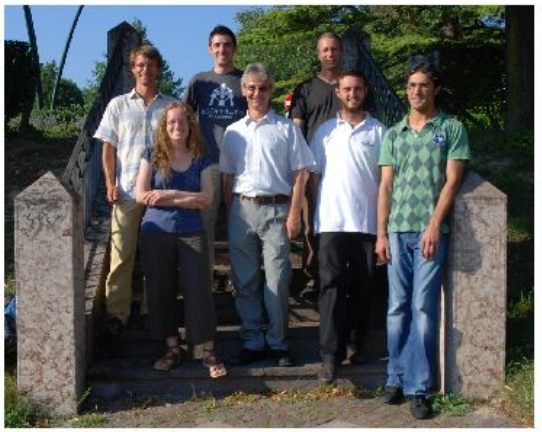
FEM GIS and Remote sensing unit:
Spatial modelling of disease vectors, biodiversity and beyond

http://gis.cri.fmach.it

People

The GIS and Remote Sensing Unit team:

- Luca Delucchi (GIS technician)
- Anne Ghisla (PhD student)
- Dr. Markus Metz (Post-Doc)
- Dr. Markus Neteler (head)
- Dr. Duccio Rocchini (Researcher)
- Dr. Roberto Zorer (Researcher)



PGIS group as of June 2011 (with Javier as guest)

Markus Neteler
Fondazione E. Mach (FEM)
Centro Ricerca e Innovazione
GIS and Remote Sensing Unit
38010 S. Michele all'Adige (Trento), Italy
http://gis.cri.fmach.it
http://www.osgeo.org
markus.neteler@fmach.it