

# Data Quality

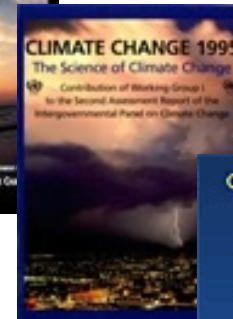
Alastair Culham & Chris Yesson

# Data sources

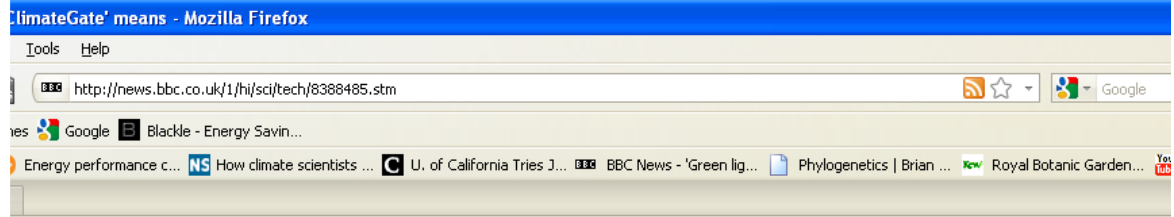
1. Climate models and their basis
2. Distribution data
3. Other data

# Climate model data

- The Intergovernmental Panel on Climate Change
- IPCC1 - 1990
- IPCC2 - 1995
- IPCC3 - 2001
- IPCC4 - 2007
- IPCC provides consensus on what scientists expect to happen
- IPCC5 is on the way



# ClimateGate



[Mobile Version](#) | [Help](#)

**BBC**

[Home](#) | [News](#) | [Sport](#) | [Weather](#) | [iPlayer](#) | [TV](#) | [Radio](#) | [More...](#)

**NEWS** **LIVE BBC NEWS CHANNEL**

[News Front Page](#)  
[World](#)  
[UK](#)  
[England](#)  
[Northern Ireland](#)  
[Scotland](#)  
[Wales](#)  
[Business](#)  
[Politics](#)  
[Health](#)  
[Education](#)  
[Science & Environment](#)  
[Technology](#)  
[Entertainment](#)  
[Also in the news](#)

[Video and Audio](#)  
[Have Your Say](#)  
[Magazine](#)  
[In Pictures](#)  
[Country Profiles](#)  
[Special Reports](#)

**Related BBC sites**  
[Sport](#)  
[Weather](#)  
[Democracy Live](#)  
[Radio 1 Newsbeat](#)  
[CBBC Newsround](#)  
[On This Day](#)  
[Editors' Blog](#)

Page last updated at 14:56 GMT, Tuesday, 1 December 2009

[E-mail this to a friend](#) [Printable version](#)

## 'Show Your Working': What 'ClimateGate' means

**VIEWPOINT**  
Mike Hulme and Jerome Ravetz

The "ClimateGate" affair - the publication of e-mails and documents hacked or leaked from one of the world's leading climate research institutions - is being intensely debated on the web. But what does it imply for climate science? Here, Mike Hulme and Jerome Ravetz say it shows that we need a more concerted effort to explain and engage the public in understanding the processes and practices of science and scientists.

As the repercussions of [ClimateGate](#) reverberate around the virtual community of global citizens, we believe it is both important and urgent to reflect on what this moment is telling us about the practice of science in the 21st Century.

In particular, what is it telling us about the social status and perceived authority of scientific claims about climate change?

We argue that the evolving practice of science in the contemporary world must be different from the classic view of disinterested - almost robotic - humans establishing objective claims to universal truth.

Climate change policies are claimed to be grounded in scientific

**THE GREEN ROOM**

A weekly series of thought-provoking opinion pieces on environmental topics

**Decision time**  
It is time for nations to end two decades of deadlock on whale conservation

[Your comments](#)

RECENT ARTICLES

- [Hooking the 'pirates'](#)
- [Whales - 'resource' or 'right'?](#)
- [Oil: Choices we need to make](#)
- [A new direction for climate policy](#)
- [Seeing REDD over forest peoples](#)
- [Reviving the spirit of Rio](#)
- [EU must increase climate ambition](#)
- [Counting the cost of alien invasions](#)
- [Has CITES had its day?](#)
- [Curious case of Kiwi hedgehogs](#)
- [Restoring natural capital](#)

LINKS

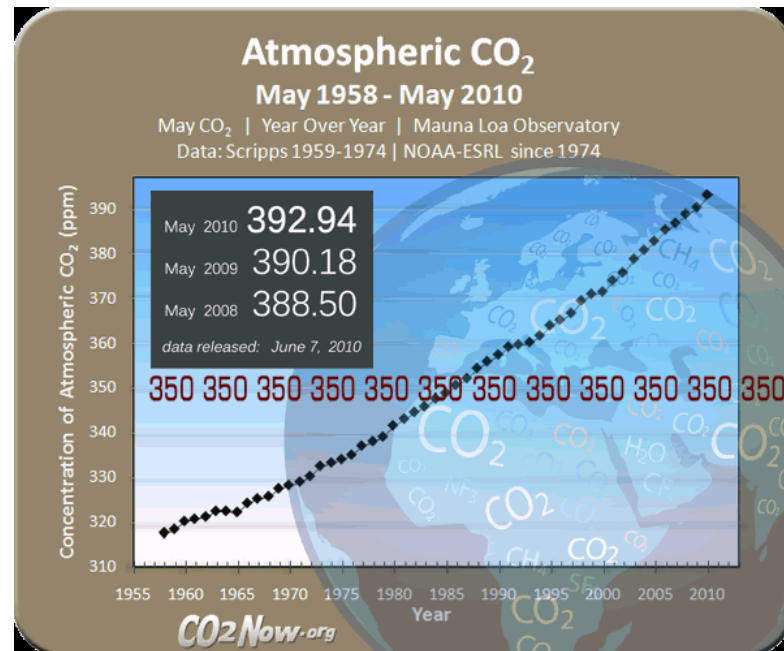
- [Copenhagen summit](#)
- [Richard Black's Earth Watch](#)
- [Earth News](#)



**“ Practising scientists know that they do not simply follow a rulebook to do their science, otherwise it could be done by a robot ”**

# Can we rely on future climate models?

- They are models, not predictions
- Sound basis in science
- Real observations – CO<sub>2</sub> Now



# How can we trust climate models for the 2080s when we don't know what the weather is tomorrow?

- Important to understand the difference between:
  - *Climate* dealing with models of general trends
  - *Weather* dealing with predictions of the specific

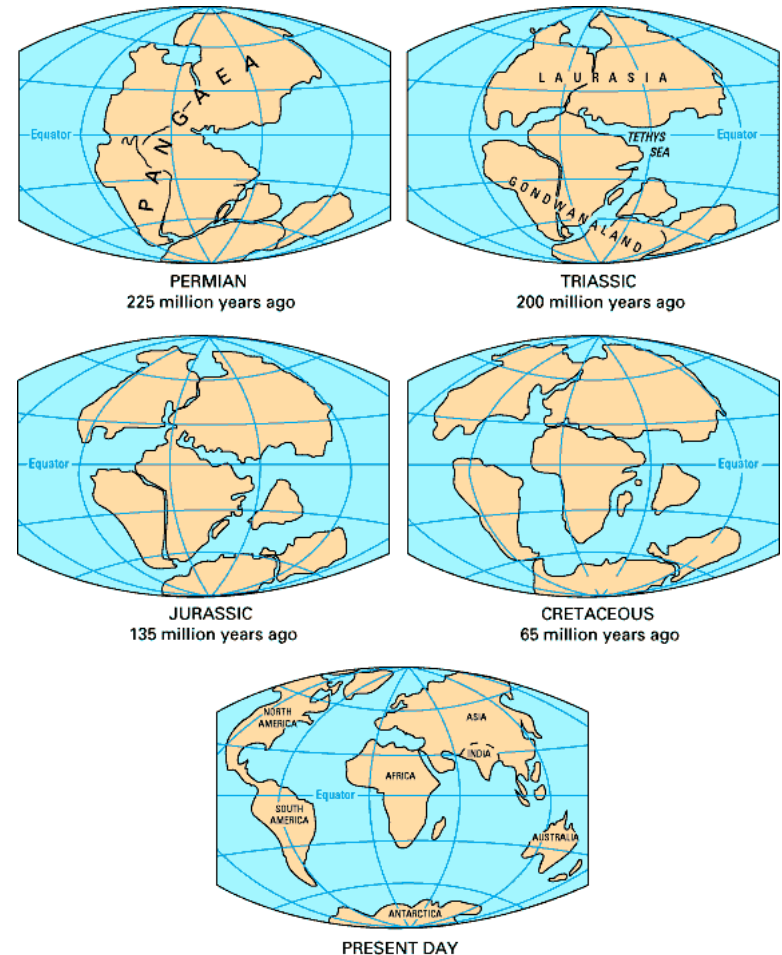
# Can we rely on past climate models?

- Modelling here relies on:
  - Knowing continental positions
  - Knowing altitudes
  - Knowing sea levels
  - Knowing atmospheric gas concentrations
- This can be validated against fossil evidence
  - Pollen/macrofossils
  - ‘Fossil’ atmospheres – from ice cores



# Palaeohistory

- Fossil history
  - Mostly pollen
- Geological record
  - Continental drift
  - Climate
- Computer models
  - Climate





# Gathering the evidence

- Fossil history is generally poor and patchy even in the best recorded groups.
- Pollen offers the best fossil record for most flowering plants.

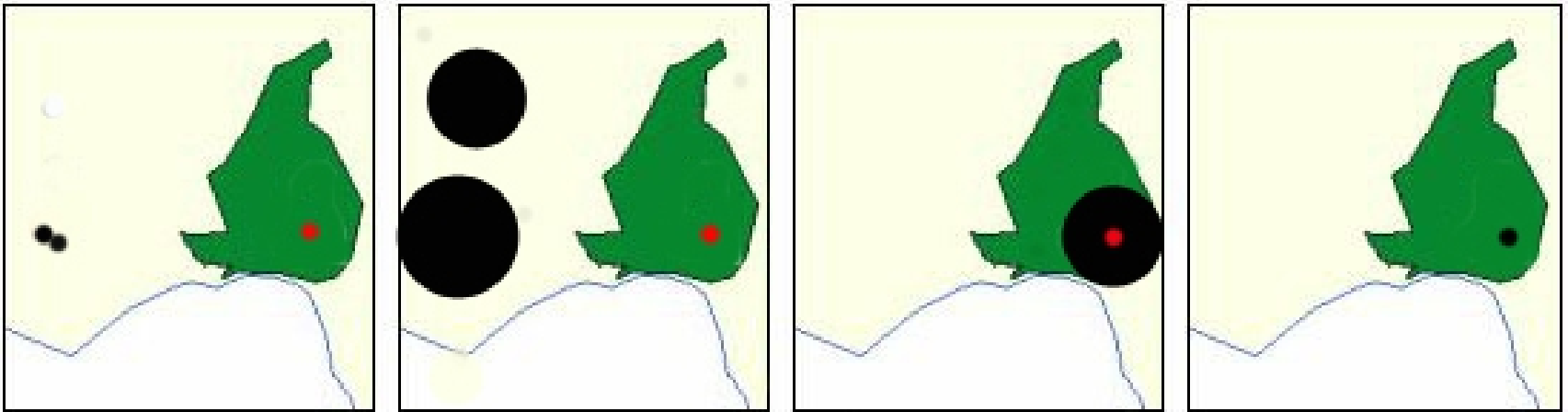


# Distribution data

- Many Sources
  - Your own validated points
  - Secondary sources
    - Individually validated
    - Batch processed

The screenshot displays the PLoS ONE website interface. At the top, there's a navigation bar with links like 'Home', 'Browse Articles', 'About', 'For Readers', 'For Authors and Reviewers', 'Journals', 'Hubs', and 'PLOS.org'. Below this, a search bar and a 'GO' button are visible. The main content area features the article title 'How Global Is the Global Biodiversity Information Facility?' by Chris Yesson et al. The article is categorized under 'RESEARCH ARTICLE' and 'OPEN ACCESS'. The abstract is partially visible, discussing the Global Biodiversity Information Facility (GBIF) and its role in providing a single point of access to species distribution data. The right sidebar contains a 'Metrics' section showing 'Total Article Views: 3233' and a 'Related Content' section with links to related subject categories and articles.

# Accuracy vs Precision



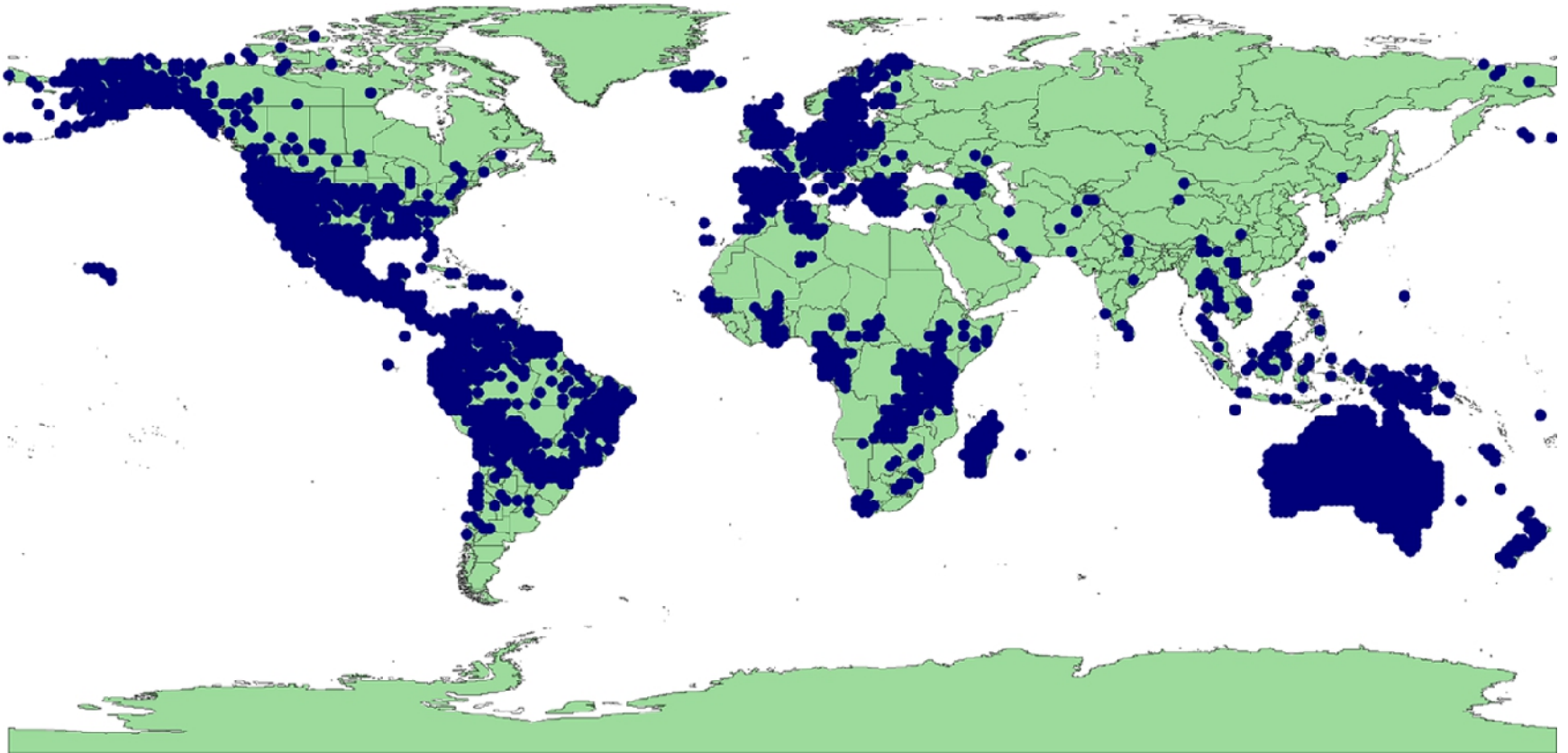
The differences between accuracy and precision in a spatial context.

The red spots show the true location, the black spots, represent the locations as reported by a collector.

- a. High precision, low accuracy.**
- b. Low precision, low accuracy showing random error.**
- c. Low precision, high accuracy.**
- d. High precision and high accuracy**

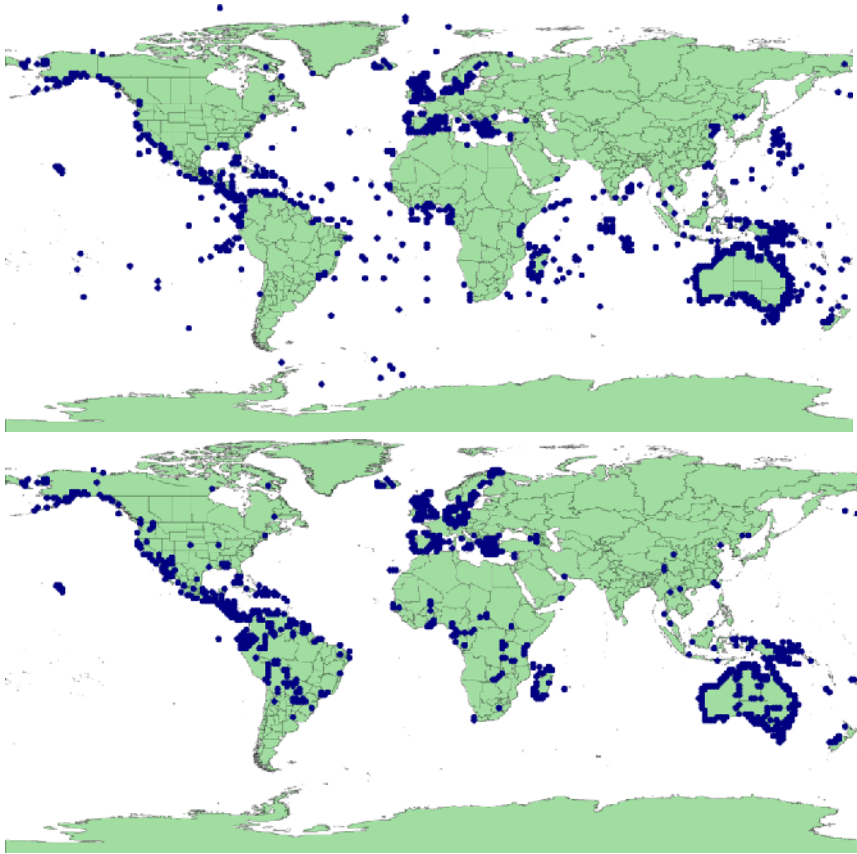
*Chapman 2005 Principles of Data Quality*

# Geographic Coverage

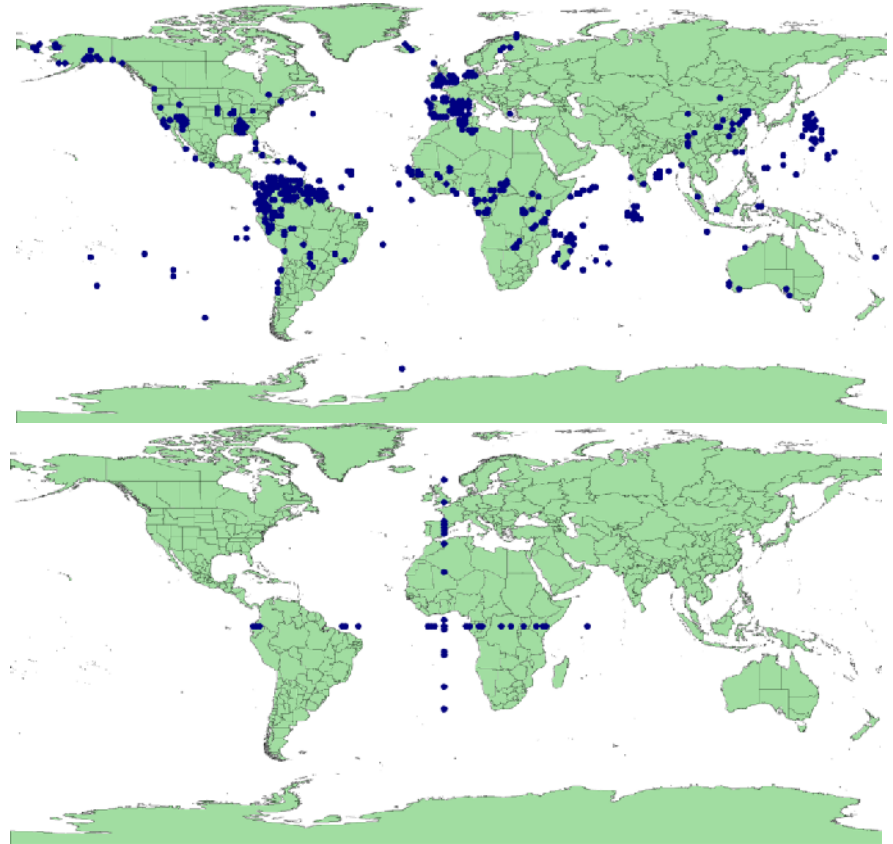


- Fabaceae data from GBIF showing patchy geographic coverage

# Basic errors

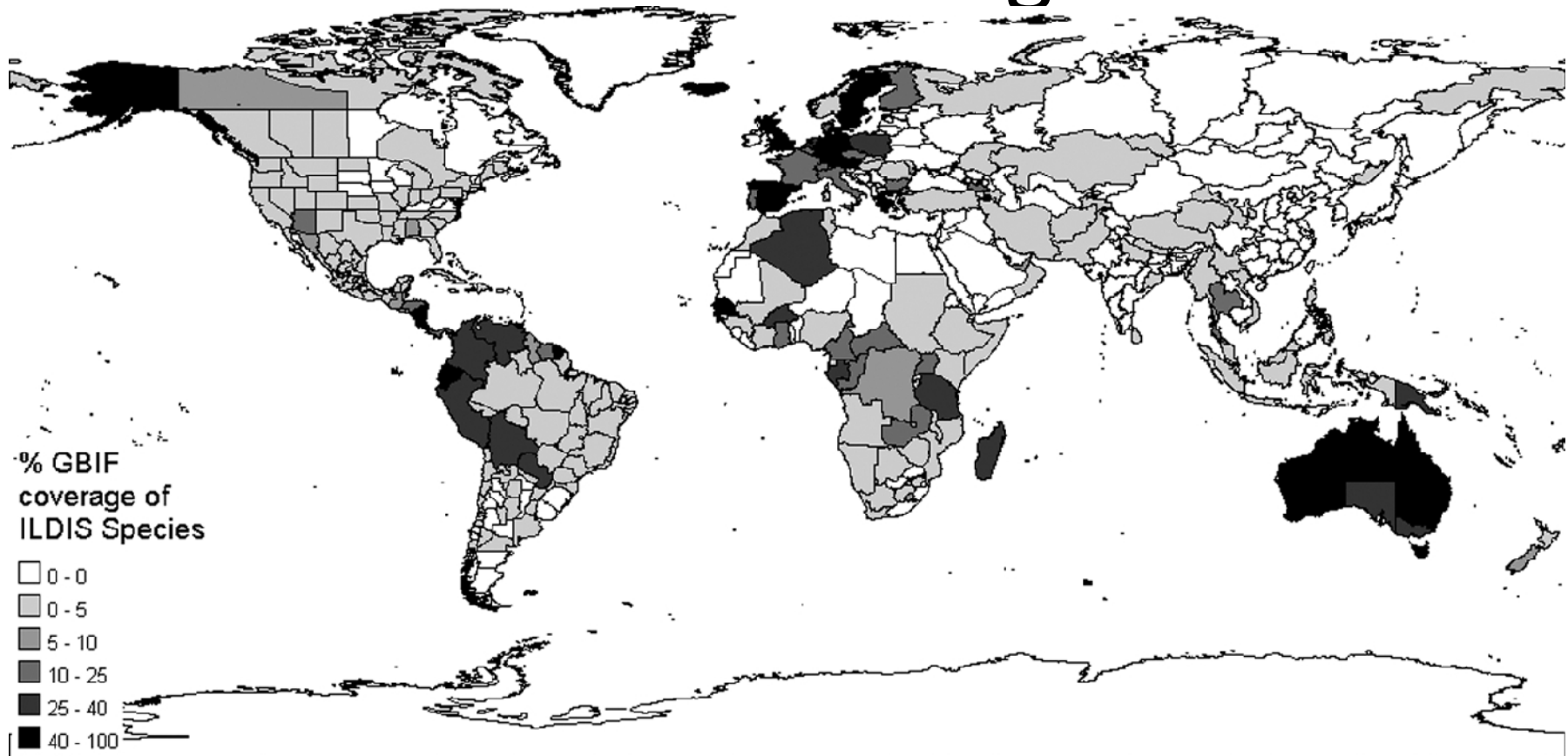


In the sea  
Near Valid



Lat/Long reversals  
Lat/Long zero

# Taxon Coverage

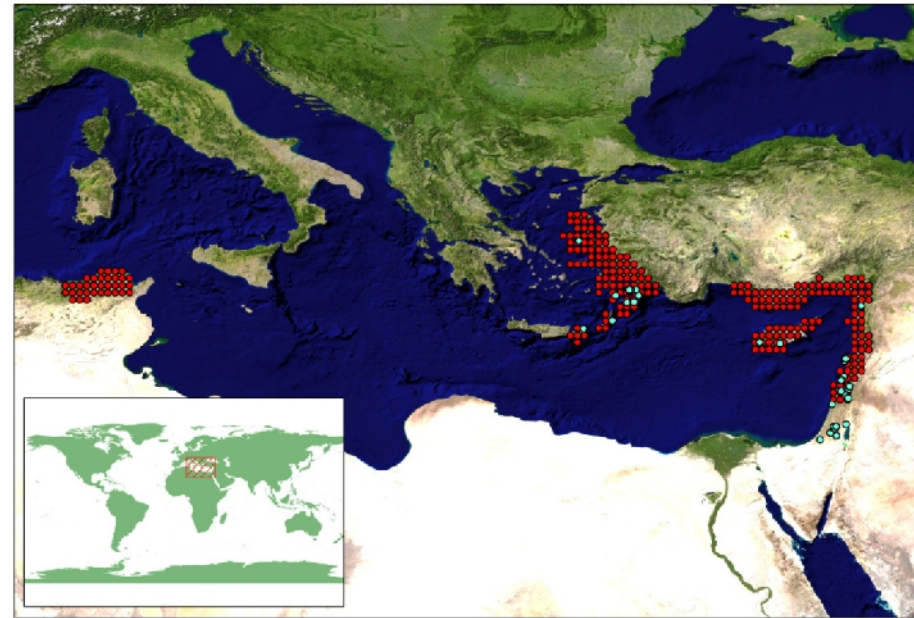
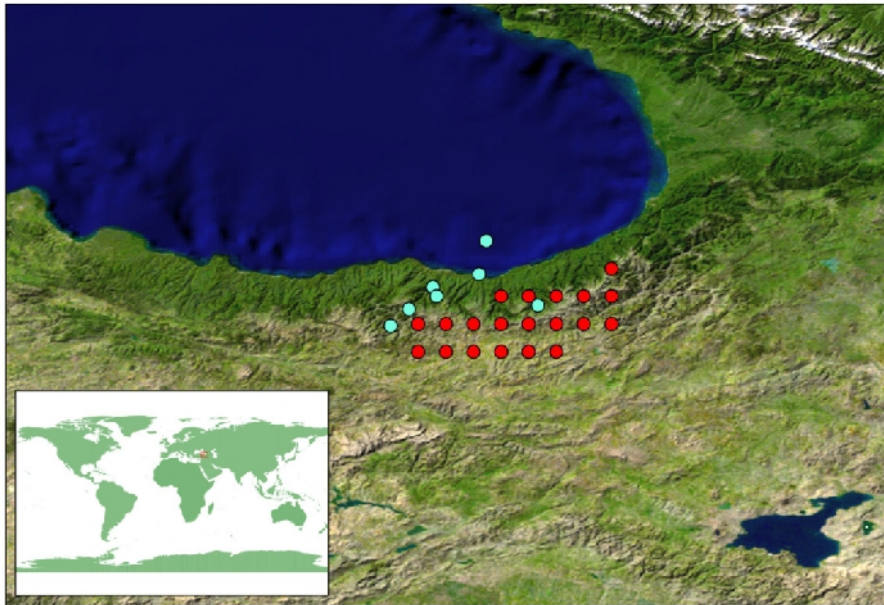


- Global Legume coverage from GBIF data per TDWG level 4 area



# Combining data

- *Cyclamen* – a data rich example
- GPS
- New localities





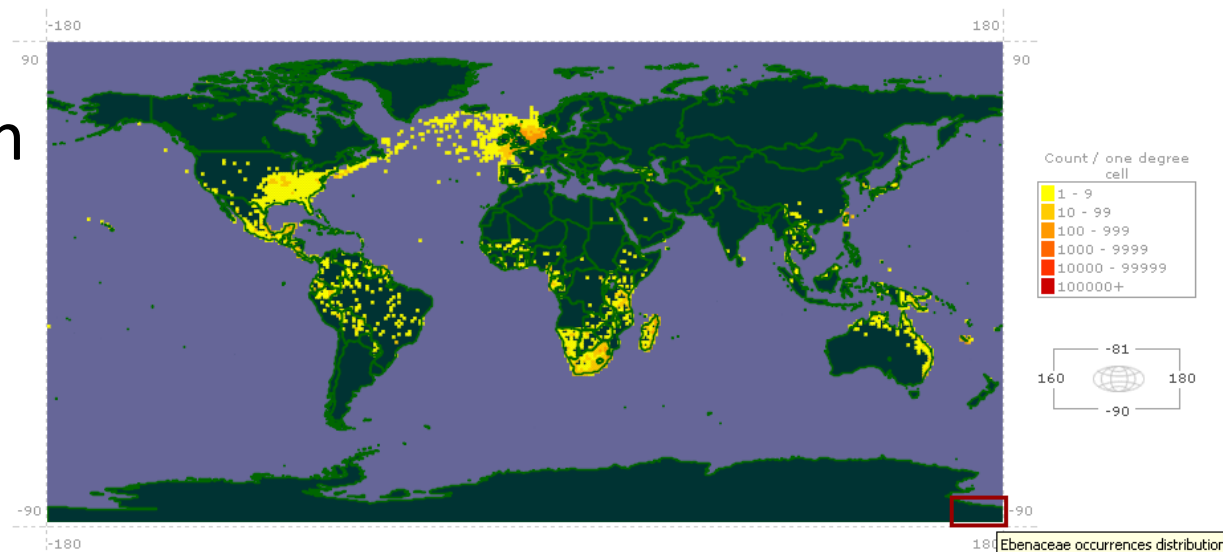
# Other issues

- Taxonomy and checklists
- Misclassified data
  - Synonymy
  - Homonymy
  - Misidentification

»Kingdom: Unknown »Family: Ebenaceae

»Kingdom: Unknown »Phylum: Angiospermae »Class: Monocotyledoneae »Order: Angiospermae »Family: Ebenaceae

## Occurrence overview



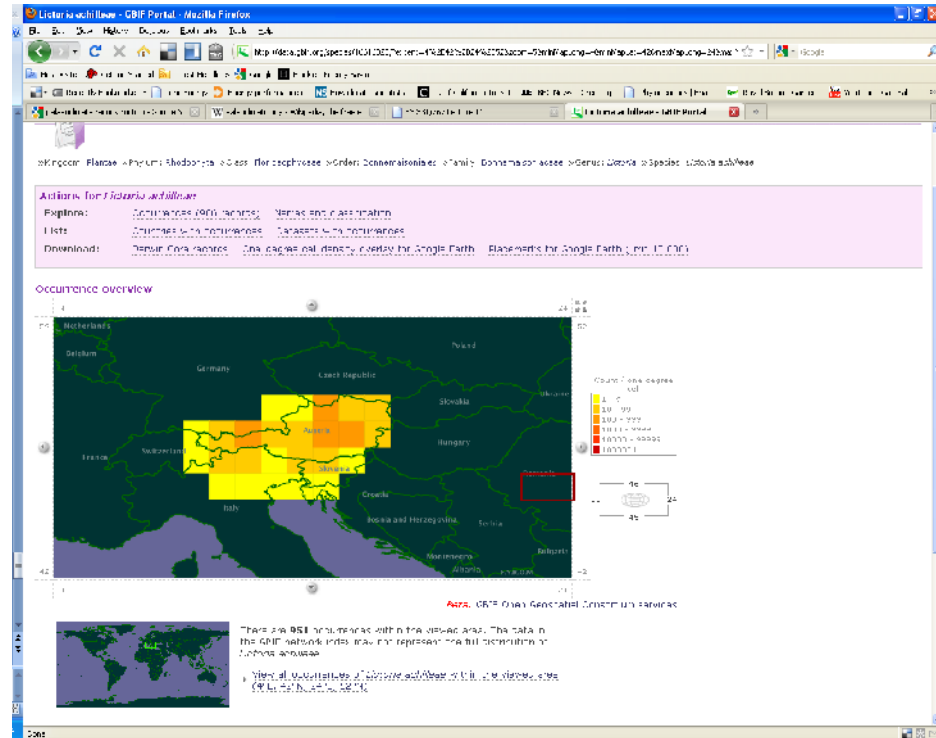
This map only shows records with coordinates (28,875 records with coordinates).

**Disclaimer:** Maps depict density of data registered within the GBIF network index and not necessarily true species occurrence density gradients. The data in the GBIF network index may not represent the full distribution of Ebenaceae.

Map includes data shared for all genera included in the family Ebenaceae (36 genera).

# More taxonomy

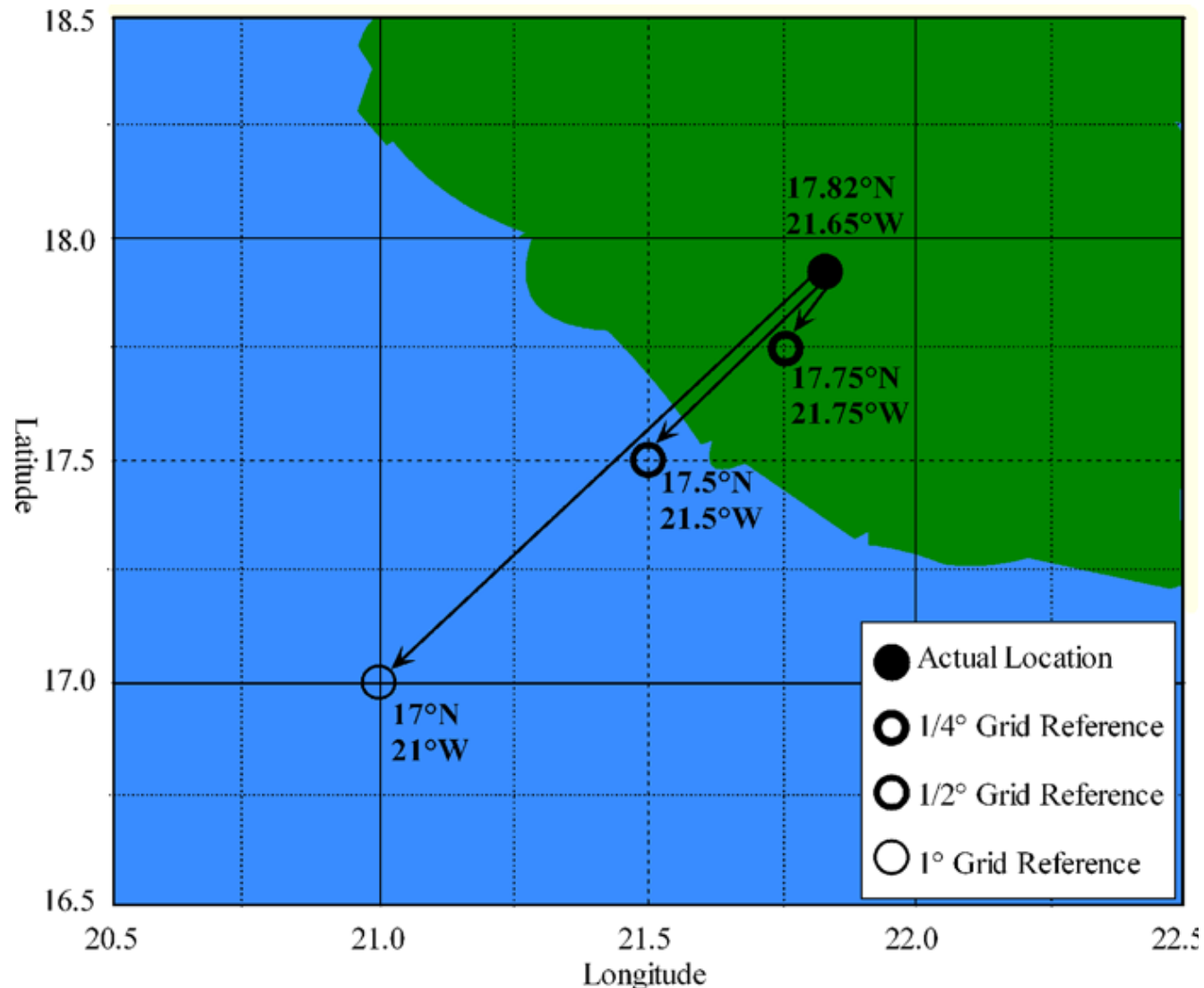
- Correctly determined taxa
- Wrongly databased
- *Lictoria achillae*
  - GBIF – listed as *Rhodophyta*
  - Source database – listed as *Lepidoptera*!



# Problems with distribution data

- Accuracy of identification
- Disambiguation
- Accuracy of record
- Context of record
- Geographical pattern to records
- <http://data.gbif.org/species/>

# Good data that appear bad



# The challenge

1. To get enough data
2. To get accurate and precise data
3. To get correctly identified records
4. To get even geographic coverage

# How Global Is the Global Biodiversity Information Facility?

- <http://www.plosone.org/article/info:doi/10.1371/journal.pone.0001124>
- Chris Yesson<sup>1</sup>, Peter W. Brewer<sup>1</sup>, Tim Sutton<sup>1</sup>, Neil Caithness<sup>1</sup>, Jaspreet S. Pahwa<sup>2</sup>, Mikhaila Burgess<sup>2</sup>, W. Alec Gray<sup>2</sup>, Richard J. White<sup>2</sup>, Andrew C. Jones<sup>2</sup>, Frank A. Bisby<sup>1</sup>, Alastair Culham<sup>1</sup>
- A review of taxonomic and geographic coverage of the GBIF database