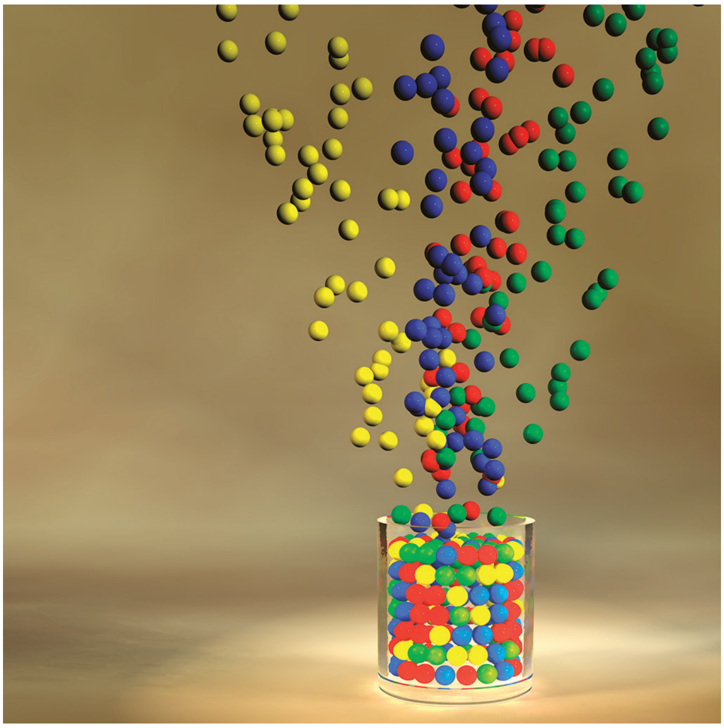
ECMWF Feature article

from Newsletter Number 116 – Summer 2008

METEOROLOGY

Implementation of TIGGE Phase I



www.ecmwf.int/en/about/news-centre/media-resources

doi:10.21957/zwwcclafbn

This article appeared in the Meteorology section of ECMWF Newsletter No. 116 – Summer 2008, pp. 10-16.

Implementation of TIGGE Phase I

Baudouin Raoult, Manuel Fuentes

The THORPEX Interactive Grand Global Ensemble (TIGGE) aims to improve the accuracy of high-impact weather forecasts by enhancing the development and use of ensemble predictions through international collaboration. The goal of TIGGE is to collect in near-real-time and store in a common format the outputs of global ensembles run to around 14 days. These outputs would be made available to researchers in operational centres and academic communities. In the longer term TIGGE would be the basis for the development of a prototype for the future Global Interactive Forecasting System (GIFS). More information about the concept of TIGGE and its objectives can be found in the article by Philippe Bougeault in this edition of the *ECMWF Newsletter* (see page 9).

The first workshop on TIGGE was held at ECMWF in March 2005. Its purpose was to collect the views of the community about the aims of the TIGGE science, the requirements for use of TIGGE data and how the associated infrastructure should be designed.

Ten operational centres volunteered to become Data Providers: these are listed in Box A on page 9 of the article by Philippe Bougeault. In addition three centres volunteered to become Archive Centres: CMA (China), ECMWF and NCAR (USA).

On 1 October 2006, three Data Providers started sending the output from their global EPS to the three Archive Centres in near-real-time. Users were provided with access to the archive on 1 May 2007. On 1 February 2008, less than three years after the first TIGGE workshop, the tenth and last of the Data Providers started sending its data in near-real-time to the Archive Centres.

This article describes the work carried out during these three years, the technical choices that have been made, and some of the difficulties that have been encountered.

Requirements

During the first workshop, the goals and requirements of TIGGE were laid out. These are now described.

A key component of TIGGE would be the TIGGE database. This would contain a core dataset consisting of ensemble forecasts generated routinely at different centres. Additional datasets could be added later to respond to requests of the scientific community, especially the THORPEX Working Groups. The database would also consist of a website providing the capability to link to regional and user-specific observational data sets.

TIGGE data would be available to all users for research purposes. Consideration would also be given to the provision of real-time access to data, in particular for demonstration projects and field experiments. The process for obtaining approval for data access would need to be transparent, streamlined and reasonably fast.

To meet user needs, a user-friendly interface to the database should be developed, that enable researchers to obtain subsets of ensemble data, especially over geographic regions of their choice.

Post-processing of archived data would be required before delivery to users. Routines would be needed for grid conversion, format conversion, and for the extraction of sub-areas, parameters and levels, for example. In addition, applications and tools for data processing, tailored to the needs of specific users, could be prepared and shared among users. These tools would need to be catalogued and documented.

Two phases were considered:

- ***Phase 1.*** The data would be collected in near-real-time (via the Internet) at a small number of central TIGGE data archives.
- ***Phase 2.*** The TIGGE database would be distributed, and a user would have to retrieve the data from the different Data Providers, using a common interface.

 doi:10.21957/zwwcclafbn

Technical implementation

This first workshop was followed by a meeting of a Working Group on Archiving held at ECMWF on 19–21 September 2005. It included representatives from CMA, ECMWF, NCAR and the North American Ensemble Forecast System (NAEFS). This group established how TIGGE Phase 1 could be implemented. The outcome of this working group was presented at the Implementation Meeting held at ECMWF on 9–10 November 2005. The participants represented both Archive Centres and Data Providers, and addressed the technical issues raised by the two preceding meetings.

Homogeneity of the TIGGE database

For the TIGGE project to succeed, it is paramount that the content of the database is as homogeneous as possible. This would ensure a productive environment with systematic data management and user access to data from many provider centres. The more consistent the archive, the easier it would be to develop applications.

All partners needed to agree on a common way of referencing data within the TIGGE dataset and describe fields using the following attributes: *analysis date, analysis time, forecast time step, origin centre, ensemble number, level* and *parameter.* In this context "parameter" refers to the physical quantity represented by the field: temperature, pressure etc.

When using fields to create a "grand ensemble", i.e. when considering all members from several Data Providers as a super ensemble, it is essential that they share the same values for the attributes *analysis date, analysis time, forecast time step, level* and *parameter.* Furthermore, all fields had to be provided using the same units. This led to the definition of the TIGGE core dataset to which all Data Providers must adhere (Table 1).

To guarantee the best precision original model grids and resolutions should be preserved if possible. Data Providers would supply interpolation routines for conversion to regular latitude-longitude grids and for point extraction. Archive Centres may endeavour to return data in regular grids using these interpolation routines.

As a common data format it was decided to use GRIB edition 2 – it is the only WMO standard that supports ensemble data without the need of local extensions. Moreover, the NAEFS community is committed to using it. Data Providers would make their data available in the archive format. Requests for clarifications and proposals for new parameters were submitted to the WMO Expert Team on Data Representation systems and Codes; as a result, a substantial number of amendments were made to the Guide to the WMO Table Driven Code Form.

The Archive Centres defined the list of GRIB2 codes, tables and templates to be used for each of the fields of the TIGGE database. They also provided guidelines (best practice) on how all TIGGE fields should be coded in GRIB2, as well as examples of properly encoded model outputs.

Data transfers

It was thought that the available network bandwidth between Europe and the USA is sufficient to meet the needs of TIGGE. However, CMA raised concerns that the current bandwidth between China and Europe as well as between China and the USA (the latter probably being better) would not meet the TIGGE requirements. Nevertheless, the situation was improved by the end of 2006 when CMA joined the GLORIAD network.

After extensive testing, it was decided that IDD/LDM (Internet Data Distribution system, Local Data Manager), an Internet based distribution system developed by UNIDATA, would suit the requirements of TIGGE for data transfer. In particular it would support the parallel transfers needed to exchange the large volumes of TIGGE data.

LDM is a broadcasting system, based on a subscription mechanism: a "downstream" LDM can subscribe to "products" from an "upstream" LDM. When a product is inserted in the upstream LDM, it is automatically sent to all the downstream LDMs that have subscribed to this product. Unfortunately, such a broadcasting system does not guarantee that products will be received by all downstream LDM, in particular if some are not running. To overcome this problem, a protocol has been defined on top of LDM to exchange fields by specifying a file name convention and a series of messages to request retransmission of missing fields. A complete description of the protocol is available on the TIGGE website.

Parameter	Unit
Parameters on a single level	
Mean sea level pressure	Pa
Surface Pressure	Pa
10 m u-velocity	m s^{-1}
10 m v-velocity	m s^{-1}
Surface temperature	K
Surface dew point temperature	K
Surface max temperature	K
Surface min temperature	K
Skin temperature	K
Soil moisture	kg m^{-3}
Soil temperature	K
Total precipitation (liquid + frozen)	kg m^{-2}
Snowfall water equivalent	kg m^{-2}
Snow depth water equivalent	kg m^{-2}
Total cloud cover	0–100%
Total column water	kg m^{-2}
Time-integrated surface latent heat flux	W m^{-2} s
Time-integrated surface sensible heat flux	W m^{-2} s
Time-integrated surface net solar radiation	W m^{-2} s
Time-integrated surface net thermal radiation	W m^{-2} s
Time-integrated outgoing long-wave radiation	W m^{-2} s
Sunshine duration	s
Convective available potential energy	J kg^{-1}
Convective inhibition	J kg^{-1}
Orography (Geopotential height at the surface)	m
Land-sea mask	0–1
Parameters on pressure levels	
Temperature	K
Geopotential height	m
U-velocity	m s^{-1}
V-velocity	m s^{-1}
Specific humidity	kg kg-1
Parameters on potential temperature surfaces	
Potential vorticity on θ = 320 K surface	K m^2 kg^{-1} s^{-1}
Parameters of potential vorticity surfaces	
Potential temperature on 2PVU surface	K
U-velocity 2PVU surface	m s^{-1}
V-velocity 2PVU surface	m s^{-1}

Table 1 List of agreed parameters and their units. Note that temperature, u-velocity, v-velocity and specific humidity are provided on the following isobaric surfaces: 1000, 925, 850, 700, 500, 300, 250 and 200 hPa. The geopotential height is provided on the same surfaces plus 50 hPa. All parameters have to be provided six-hourly. All fluxes are to be accumulated since the beginning of the forecast.

doi:10.21957/zwwcclafbn

Operational aspects

For day-to-day operations tools would have to be created to monitor the data transfer within the system. Therefore, each Archive Centre will set up a web page showing volumes, date of data and date of reception for each Data Provider. This information will be used to cross-validate the content at the three archives. In addition the Archive Centres will provide the technical coordination of the project and take on the responsibility of defining the necessary procedures.

Whenever problems arise that prevent data delivery to the Archive Centres, the Data Provider will be responsible for notifying all the Archive Centres (e.g. by sending an e-mail to the appropriate TIGGE mailing list).

The objective is to have complete data at the Archive Centres as an incomplete dataset is often difficult to use, as most of the current tools used for ensemble data assume a fixed number of members from day to day. To ensure the datasets are as complete as possible, Data Providers would endeavour to send missing data to the Archive Centres – even if this means rerunning a forecast cycle.

If an Archive Centre does not receive the expected data from a Data Provider, or if the data are incomplete or corrupted, it will first check with other Archive Centres and determine whether the failure is an isolated case. If it is an isolated case, recovery will be initiated between Archive Centres. If not, the Data Provider must re-initiate the data delivery. In any case, incidents must be investigated and documented. The Archive Centres have agreed to define and collect common metrics that can be used to create combined TIGGE-wide reports. This information will be used as a basis for the further evolutions of the system.

Participation in TIGGE must not interfere with the operational activities of Data Providers: they should be able to upgrade models, introduce higher resolutions, and make all customary changes as needed. Mechanisms should exist which allow new products from the Providers to be easily integrated into the TIGGE Archive Centres. On the other hand, Data Providers must take into account their participation in TIGGE when planning changes to their forecasting systems, and must inform Archive Centres accordingly.

To support the smooth running of TIGGE activities, it was decided that each partner would nominate two contact points: a technical contact point who will be able to address operational and technical issues, such as troubleshooting, networking or timeliness of delivery and a scientific contact point who will be able to address issues such as forecast performances or numerical errors.

User access

Access will be provided for research and education through a simple electronic registration process. Once registered, access will be given with a delay of 48 hours after the initial time of the forecast (reference time of data in GRIB2). Registration for real-time access will be handled via the THORPEX International Programme Office.

The Archive Centres will guarantee that user interfaces will present the same information (e.g. same variable names), and that similar requests, although expressed differently, should return identical results.

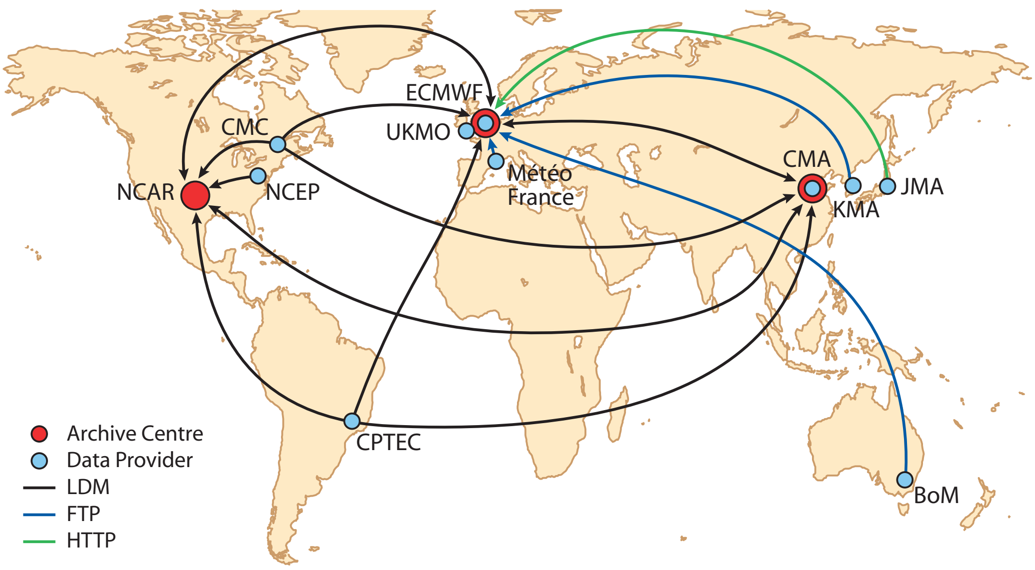


Figure 1 Exchange of data between Data Providers and Archive Centres.

doi:10.21957/zwwcclafbn

From planning to implementation

As with any project, there have been a few deviations from the original plans, mainly to accommodate the operational requirements of the Data Providers.

Collaboration

A website has been setup at http://tigge.ecmwf.int. It contains the list of all known contact points for each partner (though this information is password protected). It also contains a comprehensive description of how fields should be encoded in GRIB2, with an example of each field available for download.

A description of the protocol (file names, messages, etc.) that has been built on top of LDM can be downloaded, together with sample scripts implementing the protocol and sample configuration files.

Compliance with the agreed list of parameters

As the first data exchanges were being set up between the partners, it soon became clear that most of the Data Providers could not contribute to the whole of the agreed list of products, mainly because these products were not produced by their models. It was felt that waiting for all partners to upgrade their systems to produce the missing fields was an unnecessary delay in the building of the archive. As all the Data Providers were producing the most important fields (surface temperature, geopotential, winds, etc. on standard levels), a staged approach was chosen. Data Providers would join the project by sending whatever parameters they had, and would add more parameters during the course of the project.

Data transfer

Although LDM was the preferred solution for the exchange of data between the TIGGE partners, it was not always possible for Data Providers to install an LDM server at their site. Some decided to use either FTP or HTTP to transfer the data to one of the Archive Centres which would in turn relay it to the two others. Figure 1 shows the various protocols used between the Data Providers and Archive Centres.

Quality assurance

Good quality assurance procedures are required to guarantee the homogeneity of the dataset. Many tests have been implemented to ensure that the TIGGE database does not contain any badly encoded data that would prevent its use. With such checks in place, researchers can use the archive with confidence.

When a Data Provider starts sending new data to the Archive Centres, all new fields are marked as being in “test” mode. The new fields are checked against the agreed list of TIGGE fields, and then checked for proper encoding and units. Once all the fields have been validated, they are tagged as being in “production” mode and are sent routinely.

Although all the fields have been thoroughly validated during the test phase, changes may be introduced by the Data Providers due to evolutions in their operational environments. It is recognised that these changes may subsequently have an unforeseen effect on the TIGGE data exchange. This is why all the fields which are received in production mode are also validated, and any field that fails this validation would be quarantined for further investigation. This guarantees that no unexpected data is able to compromise the homogeneity of the archive. Figure 2 shows a screenshot of SMS, the ECMWF supervisor used to control the flow of TIGGE data at ECMWF. The “archive” task will only run once the “validate” and “quarantine” tasks are successful.

To complement the documentation and examples available from the TIGGE website, a series of GRIB2 tools were made available to all the partners. The tools included GRIB1 to GRIB2 converters and validation programmes. For example the tigge_check command will check if a field is correctly encoded and if it is part of the agreed catalogue.

Fields may be properly encoded, and pass checks by validation tools, but their values may still be incorrect. This is usually the case if the units are wrong, or the fields are instantaneous instead of accumulated. To spot these problems, a series of plots are produced every day, comparing the data from all the providers – an example of such a plot is given in Figure 3.

In order to ensure that the TIGGE archive is as complete as possible, a web page has been set up to show the status of availability of each cycle from each Data Provider. This web page (Figure 4) shows the whole history of the dataset: the addition of new fields is indicated, as well as any incidents. It is used by the partners to check the completeness of their contributions to the dataset.

 doi:10.21957/zwwcclafbn

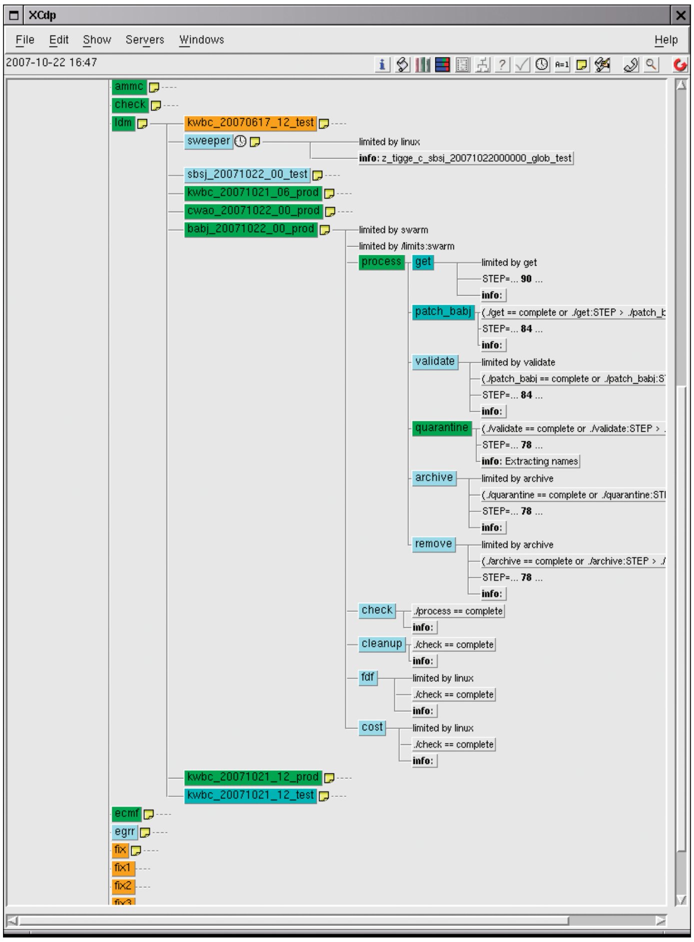


Figure 2 Processing the TIGGE flow.

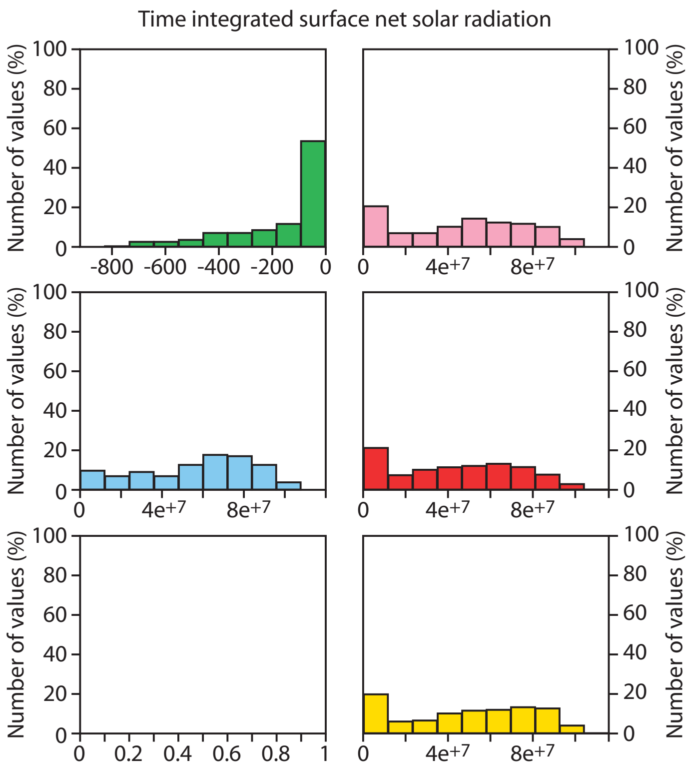


Figure 3 Example of a plot produced every day to compare data from providers. The charts of time integrated surface net solar radiation from various providers indicate an error in the data from one of the data providers (top left chart).

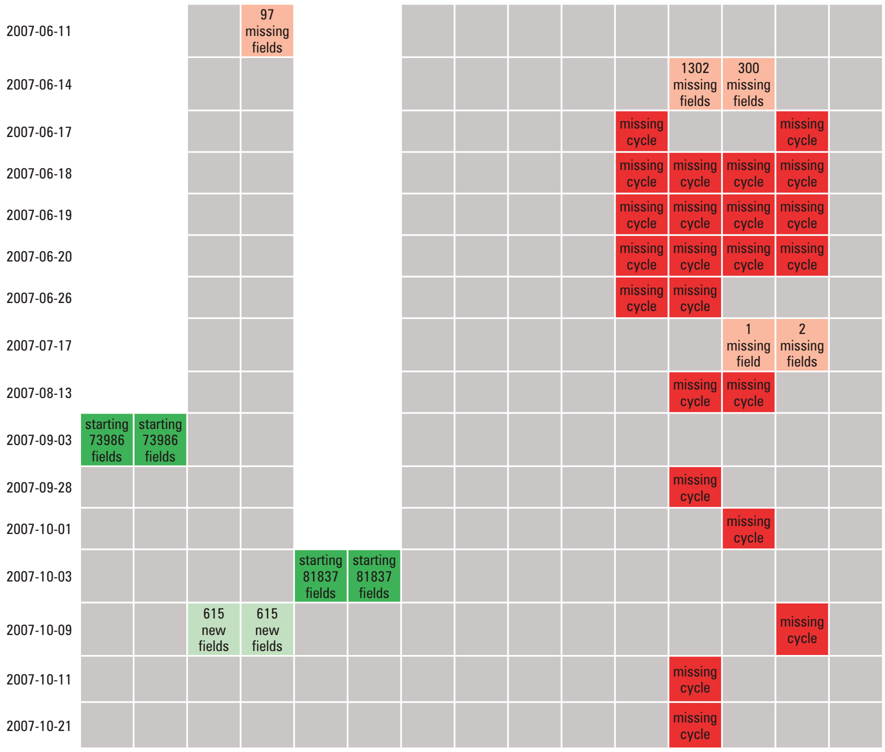


Figure 4 TIGGE history web page.

ECMWF's TIGGE portal

The ECMWF TIGGE data portal can be found at http://tigge-portal.ecmwf.int. The portal makes use of the WebMARS framework that has been developed at ECMWF to provide access to the MARS archive for external users. It allows users to select any combination of parameters, origins, levels, time steps and dates, in a very simple manner, by offering a very compact user interface and making use of the AJAX technology to control user selection. This data portal is illustrated in Figure 5.

Users have the possibility of requesting the data in various forms: on a common grid (Figure 6(a)) and on a common area (Figure 6(b)), thus retrieving a very homogeneous dataset that can easily be used. Hundreds of thousands of fields can be sought with a single request (Figure 6(c)).

As the TIGGE database is already very large, and growing, it was necessary to offer access to offline data (on tape). In order to make sure that the TIGGE activity has no adverse impact on ECMWF's core activities, TIGGE retrieval requests are handled by a dedicated SMS and a dedicated MARS server, while re-gridding is performed on a dedicated Linux server. This allows a very fine control of the resources used by this service.

There are around 100 registered users, of which a third are active. Figure 7 shows the country of origin of the registered users (excluding ECMWF internal users).

There has been a considerable growth of active users since the service started (Figure 8(a)). From December 2006 to April 2007 the service was only available inside ECMWF, but it was opened to the external users in May 2007. In addition the number of requests handled per month has increased (Figure 8(b)). These increases are reflected in the amount of data retrieved from the MARS archive (Figure 8(c)) in terms of both the amount of data retrieved from the database and the amount of data delivered to the users. The difference is due to users being given the opportunity to extract sub-areas and change the resolution of the data retrieved.

doi:10.21957/zwwcclafbn

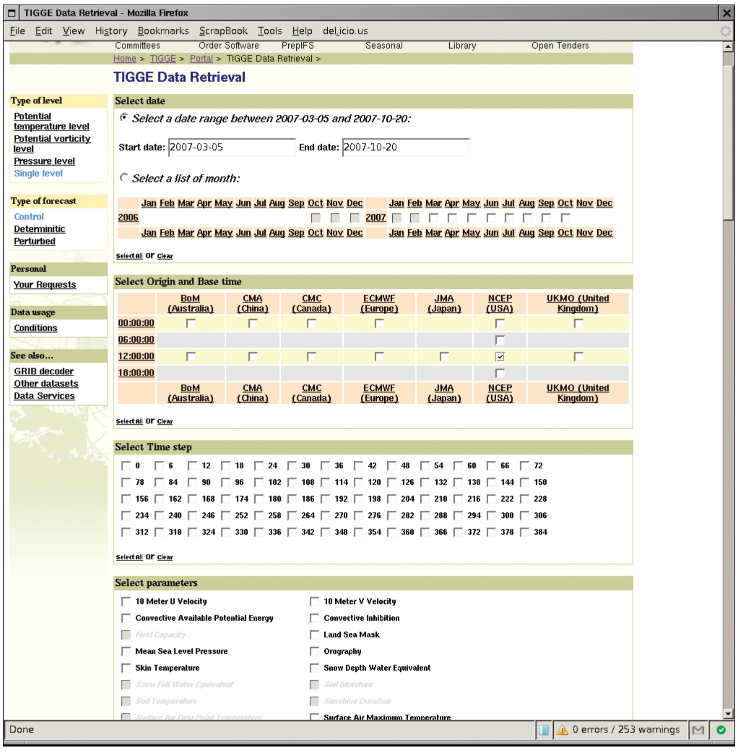


Figure 5 Selection of fields on the TIGGE data portal.

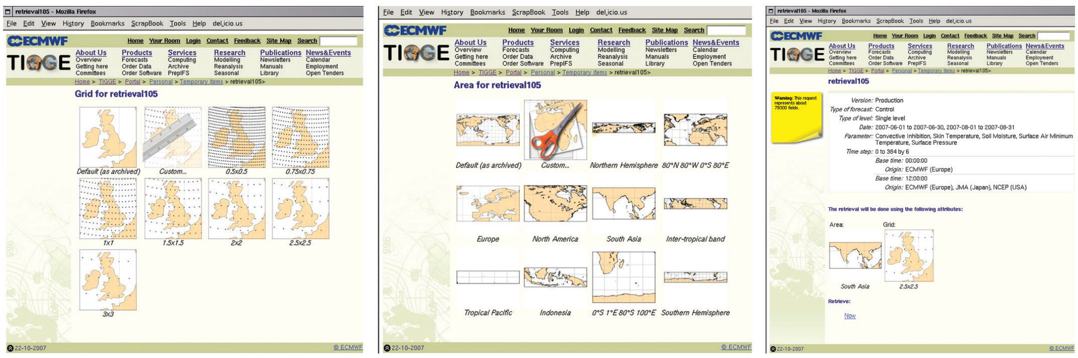


Figure 6 Web pages illustrating the selection of data in terms of (a) the grid and (b) the area and (c) the amount.

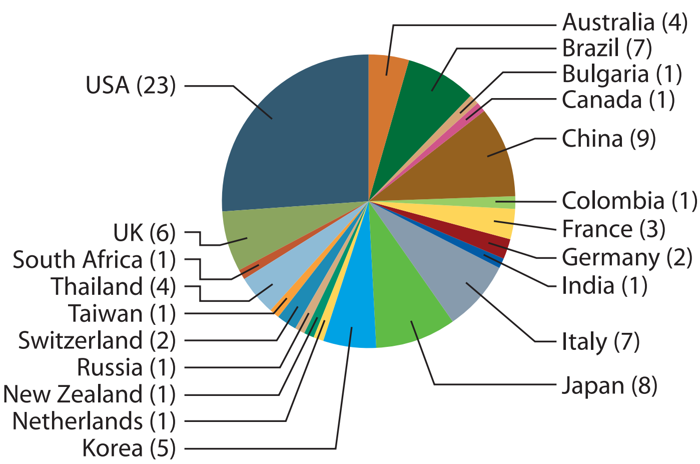


Figure 7 Number of registered users according to their country.

doi:10.21957/zwwcclafbn

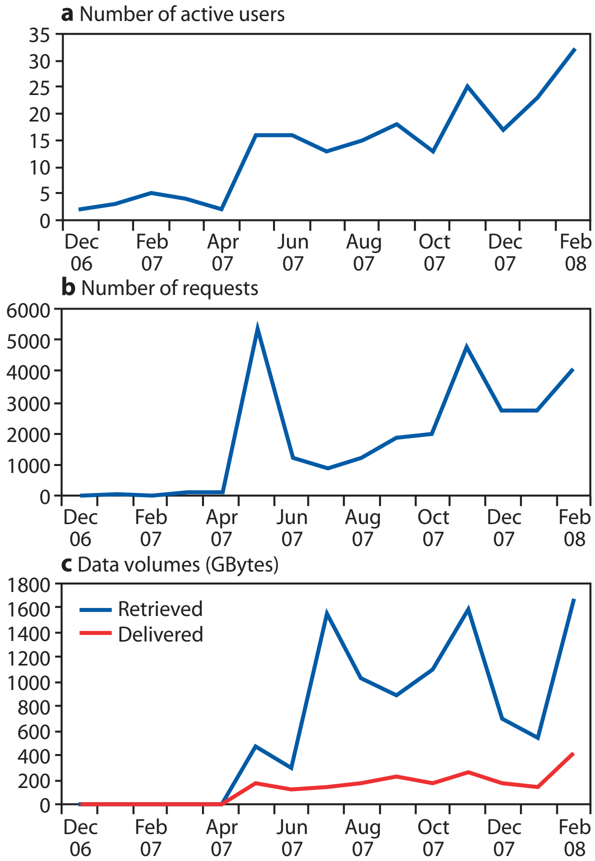


Figure 8 Variation in (a) number of active users (b) number of data requests and (c) data volumes from December 2006 to February 2008.

Success of TIGGE Phase 1

Around 240 GB (~1.6 million fields) are now exchanged routinely each day between ten Data Providers and three Archive Centres in near-real-time. The TIGGE database now contains global EPS data from all ten Data Providers, and holds more than 100 TBytes of data (600 million fields).

Around 100 users have registered with the TIGGE data portal at ECMWF, of which a third are active, generating up to 5,000 requests per month. An example of how the TIGGE data has been used is provided by the articles in this edition of the *ECMWF Newsletter* which describes some predictability studies (see page 16).

TIGGE Phase 1 is a truly successful international collaboration – it has only been possible thanks to strict governance by the Archive Centres and a strong commitment from the Data Providers. The TIGGE database now provides an essential tool for the research community, giving THORPEX the means to achieve its goals.

 doi:10.21957/zwwcclafbn