

EGU2010 SM1.3 Seismic Centers Data Acquisition session

The BRAZilian Seismographic Integrated Systems (BRASIS): infrastructure and data managementMarlon Pirchiner^{1,*}, Bruno Collaço¹, Jackson Calhau¹, Marcelo Assumpção¹, João Carlos Dourado²¹ Institute of Astronomy, Geophysics and Atmospheric Sciences (IAG), University of São Paulo (USP), Brazil² Institute of Geosciences and Exact Sciences (IGCE), State University of São Paulo (UNESP), Brazil**Article history**

Received May 5, 2010; accepted October 31, 2010.

Subject classification:

Seismology, Seismological Networks, Data Management, Earthquake Monitoring, Data Sharing.

ABSTRACT

In geophysics and seismology, raw data need to be processed to generate useful information that can be turned into knowledge by researchers. The number of sensors that are acquiring raw data is increasing rapidly. Without good data management systems, more time can be spent in querying and preparing datasets for analyses than in acquiring raw data. Also, a lot of good quality data acquired at great effort can be lost forever if they are not correctly stored. Local and international cooperation will probably be reduced, and a lot of data will never become scientific knowledge. For this reason, the Seismological Laboratory of the Institute of Astronomy, Geophysics and Atmospheric Sciences at the University of São Paulo (IAG-USP) has concentrated fully on its data management system. This report describes the efforts of the IAG-USP to set up a seismology data management system to facilitate local and international cooperation.

1. Introduction

Since the early 1990's, several temporary seismic experiments have been carried out in many areas of the Brazilian shield by the Seismological Laboratory of the Institute of Astronomy, Geophysics and Atmospheric Sciences at the University of São Paulo (IAG-USP). About 60 stations have been deployed, most of which have been operating for more than two years with broad-band sensors, for the study of the Brazilian crust and upper mantle. For many years, the large amount of data obtained was stored on various media without any correct and structured organization, making it difficult for other institutions to use these IAG-USP data. To enhance cooperation (both national and international), in 2008, we started to assemble seed volumes to be stored at the Incorporated Research Institutions for Seismology Data Management Center (IRIS-DMC).

In 2008-2009, the Brazilian oil company Petrobras approved a large project to set up a permanent Brazilian seismic network with real-time acquisition from about 60 broad-band stations. This project is being carried out by the

four main Brazilian seismology groups: the National Observatory (Rio de Janeiro), the Federal University of Rio Grande do Norte (UFRN, Natal), the Brasilia University (UnB) and the IAG-USP. In addition to continuous monitoring and the issuing of seismic bulletins, we expect to enhance seismological research in Brazil, and thus to improve our knowledge of the crustal and upper-mantle structure, the neotectonics, the crustal stress field, and the seismic hazard. Development of the human resources involved in seismology is also expected. The aspect being carried out by the IAG-USP is the project called «BRAZilian Seismographic Integrated Systems» (BRASIS).

Scientific development and productivity are increased by data sharing, scientific cooperation, education and technical training. For this reason, the acquisition, storage, management, distribution and availability of the data produced by this seismic network requires some standardization of its elementary processes to facilitate interoperability. This is achieved more easily when working with open technologies and standards [Percivall 2010].

This report describes a set of proposals and implementations to provide the Brazilian seismic network with the databases and processing systems that will best facilitate future seismological research. Here, we describe the procedures adopted by the IAG-USP Seismological Laboratory to find the best solutions to achieve these demands.

2. Compiling and sharing data from temporary deployments

Recovering data acquired over more than 15 years of temporary experiments was the first task to be undertaken. This was motivated by the data from some of the stations being gathered in the 1990's and recorded on magnetic tape that can no longer be retrieved. We regarded it to be important to ensure that our data could be safely stored and easily accessed for future use. Given the limited computational infrastructure and human resources of our laboratories, the

best solution was to store all of our previous data with the IRIS-DMC. Also, the number of access tools provided by IRIS makes our data easier to be used by the international community, increasing the scientific value of our data.

Several difficulties were experienced. Many different recording configurations and instrument types had been used even within a single station. Up to four different streams with different sampling rates and gains were used in several cases. Under such circumstances, converting a large volume of data and recompiling an equally large number of meta-information required the development of a program to help prepare the seed volumes.

The supporting materials and manuals, and the direct collaboration of the IRIS-DMC staff were essential to this task, as well as the support of the field technicians, whose organization and maintenance of field sheets served as a guide to this whole procedure. Data from about 60 stations, with half of them broad-band and most of them with one to two years of operation, are now available to the international community without restriction, through IRIS. Up until March 2010, more than 100 researchers from more than 20 countries had accessed our data.

2.1. Seed data format and tools

During this compilation, familiarization with the Standard for Exchange of Earthquake Data [SEED Manual 2010] format was necessary. Various tools used were provided by IRIS, such as converters, inspectors, selectors and verifiers of waveform data. In addition, the Portable Data Collection Center [PDCC Manual 2010] was used to manage all of the metadata.

3. Operational infrastructure

3.1. Software suite

After finishing the data compilation phase, the IAG turned its efforts to search for solutions that would meet the operational requirements of the Brazilian seismic network (the BRASIS Project). Several kinds of software were considered for acquisition and real-time processing. The requirements to facilitate cooperation between different universities and institutions in Brazil (as well as the four main institutes directly involved in the Brazilian permanent seismic network) were: low cost; open source; use of GNU/Linux operating system; possibility of local maintenance; participation in future developments; and framework platform to easily develop local tools. Two main packages were selected: Earthworm and SeisComP.

The Earthworm package was developed by the United States Geological Survey (USGS) and is now maintained by Instrumental Software Technologies Inc. (ISTI). It adopts an architecture based on the Unix philosophy, where flexibility is achieved using small specialists programs that work

together. In addition, it has an active community that discusses and solves problems. Most of the seismological centers in South America are working with Earthworm.

SeisComP was developed in the GeoForschungsZentrum (GFZ) for the GEOFON seismological network. It is structured as packages that are responsible for their functionalities that combine data acquisition, quality control, a data exchange protocol (SeedLink), an off-line data request system (ArcLink), and automatic processing and visualization tools. SeisComP was entirely redesigned in its third version [SeisComP3.org 2011], as sponsored by the German-Indonesian Tsunami Early Warning System (GITEWS) project. The software maintenance and development is now provided by gempa GmbH (global earthquake monitoring processing analysis; a private software company), while the licenses are still provided by GFZ. The main changes for this third version of SeisComP included the mediator architecture pattern, implementation of the well-accepted QuakeML data model [Schorlemmer et al. 2004], multiple Structured Query Language (SQL) database persistence (store QuakeML objects into SQL database tables), and the use of the mature open source library Spread to implement its high performance message service.

Our previous experience with the SEED format and related tools was obtained when making temporary data available for the academic community at the IRIS-DMC, and SeisComP uses the SeedLink protocol; these favored the choice of SeisComP.

SeedLink/ArcLink: RealTime/ Archived data request servers

SeedLink is a real-time data acquisition protocol based on transmission control protocol (TCP) that consists of a 512-byte Mini-SEED record and an 8-byte SeedLink header containing a sequence number. This number is used to recover data streaming in case of network errors. For SeedLink to communicate with equipment from different companies, special plug-ins are used. The SeedLink protocol is used only for real-time data streaming; it does not deal with instrument responses and does not support the full SEED.

ArcLink server is also based on TCP but it does not work with real-time data, just with requests. When ArcLink receives a request for data coming from a user, the "request handler" provides an ID that will be used to get the request status, download data, and finally delete the request. These tools are complementary and easy to install.

3.2. Hardware and network management improvements

It was necessary to review the computational structure of the laboratory, moving replicated resources in personal computers and workstations to a more dense and robust computing environment. New hardware was acquired with server architecture and virtualization support, redundant hard disks and redundant power supplies in a Gbit Ethernet network.

In this environment, different virtualized servers were set up to provide specific functionalities to seismic data management.

The efficient availability of these resources requires effective network management and monitoring, as well as routine automatic back-up. These tasks are essential to ensure a reliable service of seismological applications and analyses tools. They also add security and stability to the ongoing technological developments.

Network management and monitoring

To solve the first problem, of effective network management, a set of open-source solutions were used, which integrates applications like Nagios, Centreon, RRDTool and Cacti, to achieve the five functional areas of management network model defined by the International Organization for Standardization (ISO): Fault, Configuration, Accounting, Performance and Security (FCAPS).

Nagios incorporates features such as: network monitoring services (e.g. SMTP, POP3, HTTP, PING), host monitoring resources (e.g. processor load, disk and memory usage, running processes, log files), and environmental characteristics (e.g. temperature, relative humidity), and it allows inclusion of plug-ins to extend its monitoring capabilities, treatment of the hosts hierarchy, and escalation of staff notification according to level of warning, from different hosts and services.

Centreon uses Nagios as its core, and it offers a friendly web interface for monitoring, generating graphs of availability and service, receiving simple network management protocol (SNMP) traps, and sending alerts when a problem is detected (through SMS or email). It enables the proactive resolution of problems that involve the execution of custom scripts and commands to these alerts.

Cacti is a front-end to RRDTool. These are used together to generate graphics with a focus on the host and services performances, the delivery of service level agreement (SLA) reports of a link, and the quantification of the host or service availability. Cacti also supports SNMP for data collection on the net, and it is expandable through plug-ins like Weathermap (which generates a current network topology), Ntop (which collects statistics from the network by sniffing) and Realtime (which provides a method for displaying graphics with a resolution of up to 5 seconds).

Automated back-up routine

A back-up system has been implemented using the open source solution Bacula, a reference in enterprise environments. Bacula is a set of programs for data back-up management on multiplatform hosts, which optimizes restoration through a catalog and index of copied files. It has an independent modular structure (Director, Client, Storage, Database, Administration console) that enables distributed operation. Bacula has lots of features for customizing back-

up policies. It can be operated through the command line, but it also has some graphical interfaces. Bacula is compatible with grandfather–father–son (GFS) tape-retention strategies and allows execution of scripts before and after a job.

3.3. SeisComp3 installation

The SeisComp3 system has three major parts: Acquisition, Master (messaging communication and database access) and Trunk (automatic processing).

In the particular case of the IAG, instead of having SC3 running in the same computer, it was distributed across different virtual machines for scalability purposes. Once the hardware demands increase, because of processing needs for example, the virtual machines can be simply reconfigured and the whole SC3 system does not need to be set up again.

This type of installation enables independent control of each piece of SC3, and it also helps to better understand how they work. It is possible, for example, to fix or replace some processing modules, such as picking parameters or filter types, without having to stop the acquisition.

The installation of the IAG SC3 system was as illustrated in Figure 1:

SISLINK: the first data-acquisition server. SISLINK is responsible for acquiring external seismological data in multiple proprietary formats, running proprietary acquisition software like RTPD for RefTek equipment (such as for our station VABB) and the Apollo Server for Nanometrics (our station PMNB). Other data formats are acquired by SeedLink plug-ins, like Quanterra q330_plugin (IAG/Geoscope, SPB station), Geotech smart24_plugin (some UFRN stations) and Guralp scream_plugin (UNESP station RCLB, and UnB station JAN7). All incoming data is processed by the specific SeedLink plug-in and served as homogeneous SEED stream data.

SEISREQUEST: this catches all of the SEED data flowing from the SISLINK server, and the SC3 system starts here. SEISREQUEST runs the SeedLink archive module (slarchive) to store all data received into the SeisComp Data Structure (SDS) repository. This server also receives all of the information about the network inventories, including the related metadata, stations, stream codes, etc. All of the metadata is sent to the mediator component, to be made persistent. It runs both the SeedLink and ArcLink servers, to provide all of the acquired data and metadata for the other SC3 modules and worldwide. This module sends data from some of our stations (RCLB, VABB, PMNB) to the IRIS DMC in real time.

SEISMASTER: a server with the central mediator component (scmaster) of SC3, which is responsible for the receiving and redistribute of messages to their clients. Every

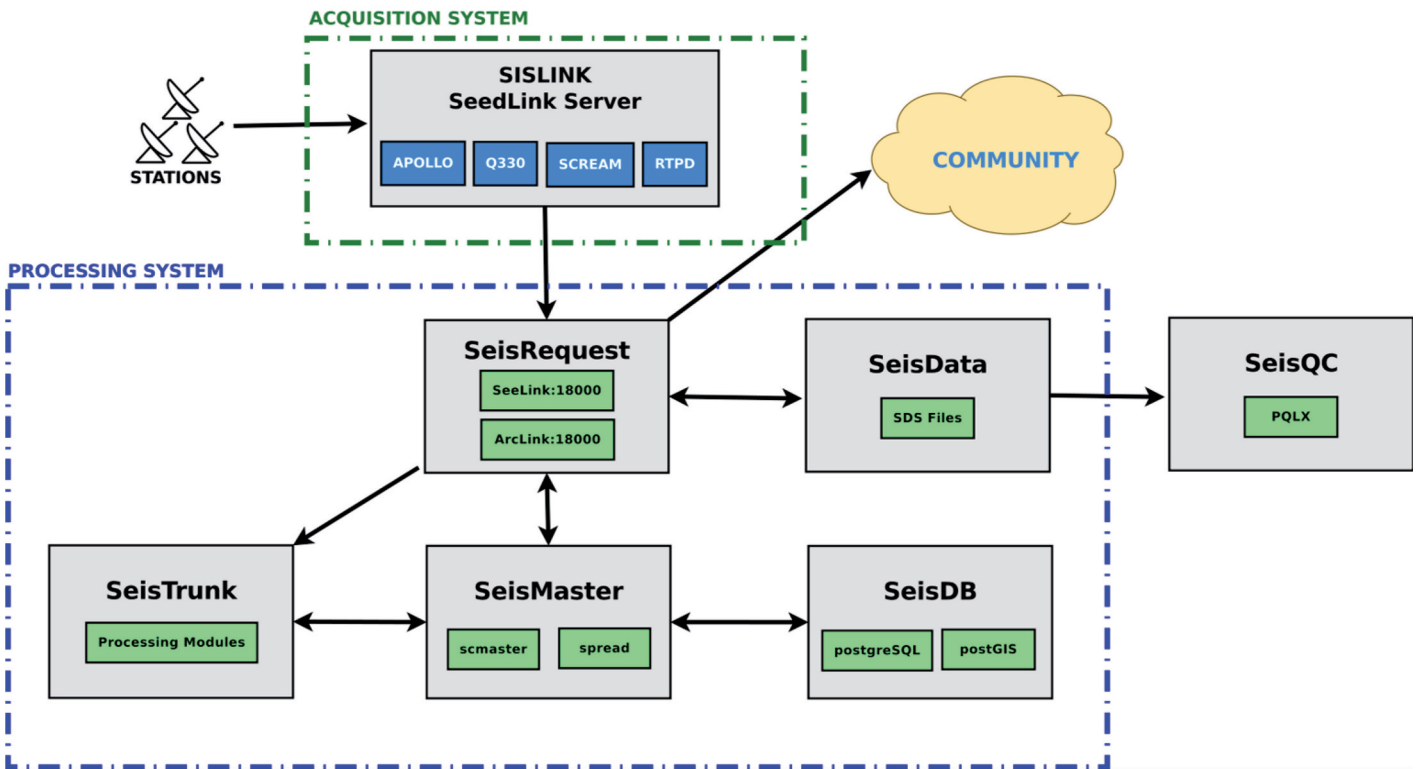


Figure 1. Schematic diagram of the SC3 acquisition and processing systems at the IAG-USP.

single module in SC3 connects to scmaster and exchanges information by this messaging system. Scmaster also persists and queries objects to the database. SeisComp3 messages are implementations of the modified QuakeML data model for real-time purposes. Each message is sent to the database before it is redistributed, increasing system integrity.

SEISTRUNK: a server that runs a set of automatic processing modules. SEISTRUNK collects quality-control metrics, does pickings, obtains amplitudes, associates picks into events and locates them. It also uses related amplitudes to provide magnitude and fast moment tensor calculations. These modules are fully customizable, so the long-term averaging/short-term averaging (LTA/STA) trigger parameters for picking can be easily changed, the finite impulse response (FIR) filters adjusted, the location algorithm changed, or the Earth model changed, for example.

SEISDB: a dedicated database server. The PostgreSQL database became the natural choice because it can be extended by PostGIS, through adding support for geographical objects and services. All of the information that SC3 writes, such as station locations, events, etc., can be used in a back-end spatial data infrastructure.

SEISGUI: runs the visual tools, like the real-time seismogram viewer (scrttv), the map with the network status (scmv), and the quality-control monitor (scqcv).

3.4. Quality control

In addition to modular deployment of SC3, all of the data acquired at the Seismological Laboratory is processed continuously by a PASSCAL quick look extended (PQLX) server that provides power spectral density (PSD) and probability density functions (PDFs) [McNamara 2004], as long-term quality-control analysis. The server portion of PQLX runs at a dedicated quality-control server (SEISQC), while the front-end is offered to all of the users by one application server.

3.5. Cooperation

To demonstrate and achieve the integration goals of this project, a basic SC3 configuration (all-in-one) was installed successfully at each of the partner institutions: UFRN (Natal), UnB (Brasilia) and the National Observatory (Rio de Janeiro). This allows each institution to acquire and share its data. In addition, each station can get real-time data from the other stations of the global networks like Global Seismic Networks (II and IU), Global Telemetry and GEOSCOPE, to form their own virtual network.

Following the IAG-USP sharing and cooperation policy, all three stations installed are freely offered by the IAG public SeedLink server. Two known clients are IRIS-DMC and the Puerto Rico Tsunami Warning Network.

3.6. Station deployment status and automatic locations

Up to November 2010, only a few stations had been installed by this project. In addition, the Brazilian seismicity

is not high enough to provide SC3 auto-location with these few stations. However, all of the equipment is ready to be deployed in 2011, when the fine tuning of the automatic locations will be made.

3.7. Archived data offered by ArcLink

Although most of the BRASIS stations are not installed yet, all of the previous data from the IAG temporary deployments are already available through ArcLink server requests. The past efforts to prepare all of the metadata and miniSEED volumes was very useful to achieve this goal. With the help of some tools provided by the SC3 package, all of the data has been rearranged in SDS and merged with real-time acquired data to be shared by ArcLink. For easier access for non-specialists, an ArcLink web interface has been set up at the Seismological Laboratory website.

4. Open-source spatial data infrastructure

Seismological studies very often require spatial data for seismo-tectonic interpretation, and they can also produce their own spatial data that can be used in future research. Stations, hypocenters, faults, cities, geological and tectonic structures, elevation models, different kinds of profiles, geophysical soundings and other data are some examples of the geospatial data related to seismology.

To help manage and provide all of this data, the Seismological Laboratory uses a dedicated server called SEISSERVICES, which implements some of the Open Source Geospatial Foundation (OSGeo) concepts and tools that offer simple spatial data infrastructure (SDI). This server runs both MapServer and GeoServer that access raster data repositories, and the PostGIS database at the SeisDB server to store and provide vector data layers.

The use of these standards enables the Seismological Laboratory to share our data in many different standard formats with different research perspectives. This also increases the possibility of the use of all of this data as services by many existing and future geographical information systems. Finally, it guarantees high interoperability at information and knowledge levels, and it will be useful for integrated data processing techniques [Percivall 2010].

5. Further perspectives

Once the real-time acquisition infrastructure is ready and running, the IAG Seismological Laboratory will concentrate its efforts on providing better on-demand seismological analysis tools.

The Seismological Laboratory traditionally uses Seismic Analysis Code (SAC) [Goldstein and Snoke 2005] and other tools for the seismic analysis and processing. A lot of time is spent in finding and preparing data to be processed by these analysis tools, as well as in converting program inputs and outputs. At the same time, it is difficult to associate and

preserve all of the parameters of the processed events to guarantee efficient use in the future.

It is important to maintain an event catalog with all of the related data as waveforms and metadata that are easily accessible by researchers and analysis tools. Open-source projects like ObsPy [Beyreuther et al. 2010], SeisHub [Barsh 2009] and SeismicHandler [Stammler 1993] will be considered in future developments of the IAG Seismological Laboratory.

6. Conclusions

All research techniques, methods and innovations depend on quality-controlled data. Many years of work can be lost in tapes or darkrooms without proper management and easy access. However, setting up an efficient data-management system is not an easy task, and a lot of resources, tools and full-time dedicated people are needed to support the research.

The SEED data format and tools provide standards for the pre-processing and sharing of seismological data. By adopting the internationally accepted common standards, it was easier for the IAG to exchange data as well as to implement newly developed analysis tools, such as the important and useful PQLX tool for quality control of station performance.

The increasing number of broad-band seismic stations and the recent telecommunication improvements imply that seismic data can be produced faster than it can be fully processed at present. For this reason, the choice of storage systems and data formats must take into account future developments in seismological analysis – the use of SEED format, QuakeML and open standards ensures the best use of our data into the future.

SeisComP3 running at the IAG-USP is performing the following tasks well: 1) acquisition of data in multiple formats; 2) archiving of data in our local database; and 3) redistribution of waveform data in seed format to other institutions. We do not anticipate difficulties in the next stages of event detection and hypocenter location, when more stations of the Brazilian network are installed.

The existence of a Spatial Data Infrastructure and Seismic Data Management Policy can promote easy and flexible ways to query data, and so should extend the productivity, creativity and innovations in research in the future.

Acknowledgments. Many thanks to all of the staff of the Brazilian seismological research centers for their close cooperation. Special thanks to Rick Benson and Mary Ann Wood from IRIS, for their help in the preparing of our first SEED volumes.

References

Barsh, R. (2009) Web-based technology for storage and processing of multi-component data in seismolog. First steps towards a new design, PhD thesis, Ludwig Maximilians Universität München, Munich, Germany, 126 pp.; available

at <http://svn.geophysik.uni-muenchen.de/trac/seishub> (last accessed 7 March 2011).

Beyreuther, M., R. Barsch, L. Krischer, T. Megies, Y. Behr and J. Wassermann (2010). ObsPy: a Python Toolbox for Seismology, *Seismol. Res. Lett.*, 81 (3), 530-533; doi: 10.1785/gssrl.81.3.530.

Goldstein, P. and A. Snoke (2005). SAC Availability for the IRIS Community, Incorporated Research Institutions for Seismology Newsletter, April 7, 2005, UCRL-JRNL-211140, 6 pp.

McNamara, D.E. and R.P. Buland (2004). Ambient Noise Levels in the Continental United States, *B. Seismol. Soc. Am.*, 94 (4), 1517-1527.

PDCC Manual (2010). Portable Data Collection Center (PDCC) v3.7 User Manual, by Robert Casey, IRIS DMC, August 2010, 53 pp.; http://www.iris.edu/pub/programs/pdcc/PDCC_3.7_User_Manual.pdf.

Percivall, G. (2010). The application of open standards to enhance the interoperability of geoscience information, *International Journal of Digital Earth*, 3 (S1), 14-30.

Schorlemmer, D., A. Wyss, S. Maraini, S. Wiemer and M. Baer (2004). QuakeML - an XML schema for seismology, *ORFEUS Newsletter*, 6 (2); <http://www.orfeus-eu.org/Organiza-tion/Newsletter/vol6no2/quakeml.shtml> (last accessed 7 March 2011).

SEED Manual (2010). SEED Reference Manual, SEED Format Version 2.4, May 2010, IRIS, 212 pp.; http://www.iris.edu/manuals/SEEDManual_V2.4.pdf.

SeisComp3.org (2011). SeisComp 3 documentation; <http://www.seiscomp3.org/wiki/doc> (last accessed 7 March 2011).

Stammler, K. (1993). Seismichandler—Programmable multichannel data handler for interactive and automatic processing of seismological analyses, *Computers & Geosciences*, 19 (2), 135-140; doi: 10.1016/0098-3004(93)90110-Q.

Websites

Bacula: <http://www.bacula.org>
 Cacti: <http://www.cacti.net>
 Centreon: <http://www.centreon.com>
 EarthWorm: <http://folkworm.ceri.memphis.edu/ew-doc>
 GeoNetwork: <http://geonetwork-opensource.org/>
 GeoServer: <http://geoserver.org>
 GFS: http://en.wikipedia.org/wiki/Grandfather-father-son_backup
 MapServer: <http://mapserver.org>
 Nagios: <http://www.nagios.org>
 Ntop: <http://www.ntop.org>
 OGC: <http://www.opengeospatial.org>
 OSGEO: <http://www.osgeo.org>
 PDCC: <http://www.iris.edu/software/pdcc>
 PostGIS: <http://postgis.refractor.net/>
 PQLX: <http://www.iris.edu/software/pqlx>

Realtime: <http://docs.cacti.net/plugin:realtime>
 RRDtool: <http://www.mrtg.org/rrdtool>
 SeisComp3: <http://www.seiscomp3.org>
 SLA: <http://www.service-level-agreement.net>
 Spread: <http://www.spread.org>
 WeatherMap: <http://netmon.grnet.gr/weathermap>.

*Corresponding author: Marlon Pirchiner,
 Institute of Astronomy, Geophysics and Atmospheric Sciences (IAG),
 University of São Paulo (USP), Brazil; e-mail: marlon@iag.usp.br.

© 2011 by the Istituto Nazionale di Geofisica e Vulcanologia. All rights reserved.