7.6

DEVELOPMENT OF CEOP REFERENCE SITE DATA SETS

Scot M. Loehrer*, Linda E. Cully, and Steven F. Williams

Joint Office for Science Support
University Corporation for Atmospheric Research, Boulder, CO

1. INTRODUCTION

The Coordinated Enhanced Observing Period (CEOP) was advanced by the World Climate Research Programme (WCRP) as a way to integrate the research and data activities of the eight Global Energy and Water Cycle Experiment (GEWEX) Continental Scale Experiments (CSEs) and other associated experiments (Fig. 1) (Koike, 2004). The science questions being examined as part of CEOP include documenting and assessing the structure of global monsoon systems and better documentation and simulation of the water and energy budgets at all time scales. In order to develop a better understanding of these science questions, CEOP developed a plan to develop and archive the needed suite of data sets including in-situ and satellite observations and model output focused on a number of locations around the world termed reference sites (CEOP 2001). Each of the CSEs selected a number of well-instrumented locations within their particular region to serve as reference sites. A total of 36 reference sites were chosen covering various climatic regimes around the globe, everything from the arctic to the tropics and from below sea level to 5000 m above sea level (Fig 2). The inclusion of a number of reference sites in four of the worlds monsoon regions as well as in the far northern latitudes lead to the involvement of the WCRP Climate Variations and Predictability (CLIVAR) and Climate and Cryosphere (CliC) programs.

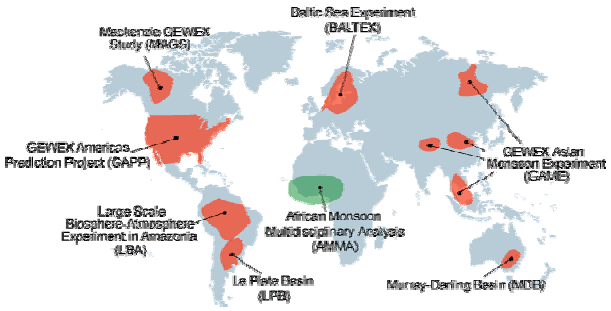

Figure 1. Locations of GEWEX Continental Scale Experiments.

*Corresponding Author Address: Scot Loehrer, UCAR/JOSS, P.O. Box 3000, Boulder, CO 80307-3000; email: loehrer@ucar.edu

The reference sites provide a varying set of observations including surface meteorology, radiation, fluxes, soils, and atmospheric profiles in a variety of temporal resolutions and in a number of formats. In order to make these disparate reference site data more readily usable and accessible to the scientific community CEOP determined that standard formats, parameter names, units, temporal resolution and quality assurance methodologies should be used. The CEOP Data Archive (CDA) at the University Corporation for Atmospheric Research/Joint Office for Science Support (UCAR/JOSS) conducts these tasks for CEOP as well as providing for the archival and dissemination of the final CEOP reference site data sets. This extended abstract summarizes the data sets developed for CEOP and the methodologies utilized in the processing and quality assurance of these data as well as summarizing the current status of the archive.

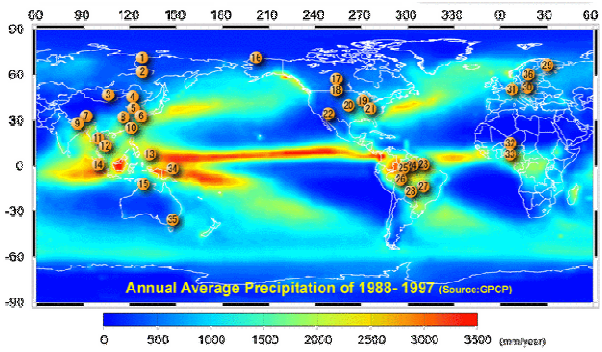

Figure 2. Locations of CEOP reference sites.

2. CEOP ENHANCED OBSERVING PERIODS

CEOP is comprised of four data collection periods termed Enhanced Observing Periods (EOPs) during which every operational reference site is to provide data to the archive (Fig. 3). EOP-1 (1 July to 30 September 2001) was a preliminary data period in which a subset of reference sites submitted data to the archive and the procedures used to create the reference site data sets were developed, tested and examined. EOP-2 (1 October 2001 to 30 September 2002) was a buildup phase during which the supporting data sets from the new satellites and model products were being phased in. The lessons learned during the development of the EOP-1 reference site data sets were used to develop new procedures for the two annual cycle data sets, EOP-3 (1 October 2002 to 30 September 2003) and EOP-4 (1 October 2003 to 31 December 2004). The

following sections provide details on the techniques used to develop the CEOP reference site data sets.

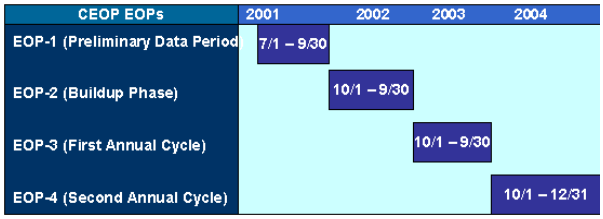

Figure 3. Schedule of CEOP Enhanced Observing Periods.

3. CEOP EOP-1 DATA SET DEVELOPMENT

As mentioned in the previous section, EOP-1 was a preliminary data period for the testing and examination of methodologies to create the reference site data sets. A prototype reference site data set was developed by the CDA for review by the CEOP community to determine its utility for answering the CEOP science questions. For EOP-1 two types of data sets were developed for each reference site that submitted data, 1) an hourly resolution surface meteorology, flux and radiation data set (Table 1) and; 2) an hourly resolution soils data set (Table 2). A total of six CSEs submitted data from 16 of the reference sites for EOP-1. The data and accompanying documentation were submitted in their raw formats and at varying temporal resolutions. Figure 4 contains an overview of the process the CDA developed to create the EOP-1 reference site data sets.

Table 1. Parameters included in the CEOP EOP-1 Hourly Surface Meteorology, Flux and Radiation Data Sets.

Parameter	Units
Station Pressure	hPa
Air Temperature	Degrees C
Dew Point	Degrees C
Relative Humidity	%
Specific Humidity	g/kg
Wind Speed	m/s
Wind Direction	Degrees
U Wind Component	m/s
V Wind Component	m/s
Total Precipitation	mm
Snow Depth	cm
Sensible Heat Flux	W/m^2
Latent Heat Flux	W/m^2
Incoming Shortwave	W/m^2
Outgoing Shortwave	W/m^2
Incoming Longwave	W/m^2
Outgoing Longwave	W/m^2
Net Radiation	W/m^2
Skin Temperature	Degrees C
CO_2 Flux	$\mu Moles/m^2/s$
Incoming PAR	$\mu Moles/m^2/s$
Outgoing PAR	$\mu Moles/m^2/s$

Table 2. Parameters included in the CEOP EOP-1 Hourly Soils Data Sets.

Parameter	Units
Soil Heat Flux	W/m^2
Soil Temperature	Degrees C
Volumetric Soil Water Content	%

The raw data from each site were converted into common columnar ASCII formats for each of the two final data set types and higher resolution data were put into the common temporal resolution of 60 minutes. The converted data sets were then passed through a set of gross limit checks (similar to those shown in Table 8) to ensure that the values of each parameter were reasonable for a location on the surface of the Earth. Each parameter from each station to be included in the final data set was then plotted and visually examined to detect more subtle problems. Any issues found in any part of this process were noted and returned to the data source for verification and/or correction. There are no data quality flags provided as part of the EOP-1 data sets. Finally, once any issues have been resolved, the data sets were made available to the scientific community.

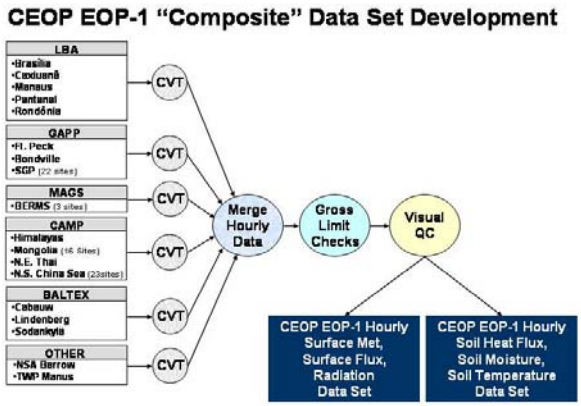

Figure 4. EOP-1 data set development process.

4. CEOP EOP-3 DATA SET DEVELOPMENT

The utility of the EOP-1 data sets for furthering progress on the CEOP science questions and improvements for the annual cycle data sets were some of the issues examined as part of several model-data intercomparison studies (e.g. Lu and Mitchell 2004 and Lidard, et al. 2004). The CEOP community gathered in Berlin, Germany in March of 2003 for the Second CEOP Implementation Meeting and the CEOP Reference Site Managers Workshop to determine the direction for the development of the subsequent two annual cycle data sets.

Several significant changes were recommended for the development of the annual cycle data sets. First among these was the need to have 30-min resolution

data from all sites to get a more complete representation of the atmospheric budgets and to better coincide with the polar orbiting satellite overpasses for intercomparisons. Second, it was recommended that meteorological tower data be included in the annual cycle data sets so that the lower levels of the atmosphere could be examined at a higher temporal resolution than allowed by 12-hourly radiosonde releases.

Another recommendation was to define two categories of data to be submitted to the CDA. Category 1 included common or low exploitation value data where the measurement technology was common and generally well understood. These data were to be submitted to the CDA within 6 months of collection. Category 2 included high exploitation value data where the measurement technology was more sophisticated or experimental in nature. Given the nature and complexity of these data, they were to be submitted to the CDA within 15 months after collection. The individual data providers determined which data at their sites were in which category. Typically, the Category 2 data included the flux parameters and therefore it was recommended that the flux data be a separate data set so as not to delay the release of other parameters. Thus the annual cycle data sets were comprised of up to four data sets at each reference site: 1) a surface meteorology and radiation data set (Table 3), 2) a meteorological tower data set (Table 4), 3) a soils data set (Table 5) and; 4) a flux data set (Table 6).

Table 3. Parameters included in the annual cycle Surface Meteorology and Radiation Data Sets.

Parameter	Units
Station Pressure	hPa
Air Temperature	Degrees C
Dew Point	Degrees C
Relative Humidity	%
Specific Humidity	g/kg
Wind Speed	m/s
Wind Direction	Degrees
U Wind Component	m/s
V Wind Component	m/s
Total Precipitation	mm
Snow Depth	cm
Incoming Shortwave	W/m^2
Outgoing Shortwave	W/m^2
Incoming Longwave	W/m^2
Outgoing Longwave	W/m^2
Net Radiation	W/m^2
Skin Temperature	Degrees C
Incoming PAR	$\mu Moles/m^2/s$
Outgoing PAR	$\mu Moles/m^2/s$

Table 4. Parameters included in the annual cycle Meteorological Tower Data Sets.

Parameter	Units
Sensor Height	m
Station Pressure	hPa
Air Temperature	Degrees C
Dew Point	Degrees C
Relative Humidity	%
Specific Humidity	g/kg
Wind Speed	m/s
Wind Direction	Degrees
U Wind Component	m/s
V Wind Component	m/s

Table 5. Parameters included in the annual cycle Soil Temperature and Moisture Data Sets.

Parameter	Units
Sensor Height	m
Soil Temperature	Degrees C
Volumetric Soil Water Content	%

Table 6. Parameters in the annual cycle Flux data sets.

Parameter	Units
Sensor Height	m
Sensible Heat Flux	W/m^2
Latent Heat Flux	W/m^2
CO_2 Flux	$\mu Moles/m^2/s$
Soil Heat Flux	W/m^2

Another recommendation was data quality flags be included for each parameter in the data set. The data quality flags were set by each reference site based on their chosen methodology. Even though this does not provide a consistent quality assurance methodology throughout the data set, it allows those most familiar with the data to determine data quality. Descriptions of the methods used and often reasons for specific flags being applied were included as part of the documentation files submitted by each reference site. The data quality flags in the data along with the summaries included in the documentation provide guidance to the community on how to best utilize the data.

A final recommendation was that each CSE convert the raw data from their reference sites to a set of common columnar ASCII formats and these converted data were then submitted to the CDA. To provide guidance to the CSEs in this process, the CDA developed a Reference Site Data Set Procedures Report (CEOP 2004) which described in detail each of the four data formats including parameter names, units, the structure of each data record, missing values to be

used, and even some equations were provided for the derivation of selected parameters. In addition, the report described in detail the information to be included in the documentation file sent with the data files. This report was approved by the CEOP Scientific Steering Committee and was the guiding document for the development of the annual cycle data sets.

After taking these recommendations into account the CDA developed a new set of procedures for the development of the annual cycle data sets as shown in Fig. 5. Since the data sets were converted to the set of common formats by the reference sites themselves, the CDA placed substantial effort in the development of highly detailed quality assurance software to ensure that the format conversions were properly completed. Additionally, the CDA was able to use its experience in developing the EOP-1 data sets to develop an additional series of automated checks to ensure the data were as complete and accurate as possible.

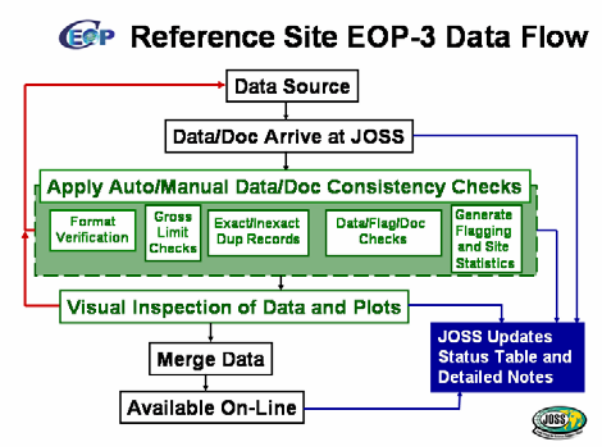

Figure 5. EOP-3 data set development process.

The automated checks focused on ensuring the integrity of the data format are summarized in Table 7. The gross limit checks conducted on each parameter in the annual cycle data sets are shown in Table 8. Between the detailed format checks and the gross limit checks many of the most common and severe data set problems were discovered. These included data submitted in the incorrect format (to the level of ensuring decimal points, spaces and colons were placed correctly), improperly applied data quality flags, incorrect derivation of parameters, parameters and flags placed in improper columns, incorrect units, improper sensor heights, changing station locations, among others. In addition to these automated checks, a number of statistics are produced for each data set that can also highlight potential problems. These include extreme values for each parameter, counts and percent of each flag type for each parameter at each height, listings of all station locations found, listing of all sensor heights found, and summaries of total error counts for the data at each height and the metadata.

Table 7. Some of the automated checks conducted by the CDA on the annual cycle data sets.

1) File has proper temporal sorting.
2) Duplicate record check.
3) Decimal points, spaces, colons in proper locations.
4) In-exact duplicate record check (same time reported but different data values)
5) Correct record length.
6) No control characters.
7) Verify metadata and data field location and justification.
8) Consistency between file name and metadata and data within the file.
9) Consistency between nominal and actual times.
10) Gross limit checks on every metadata and data value (see Table 8).
11) Valid sensor heights.
12) Valid data quality flag values.
13) Missing data value flagged as missing.
14) Data flagged as missing has the proper missing data value.
15) Identify completely missing records (i.e. data gaps).
16) Station location constant for all data records.

Table 8. Gross limit checks utilized by the CDA for the annual cycle data sets.

Parameter	Max Value	Min Value
Air Temperature	-90°C	60°C
Dew Point	-90°C	35°C
Relative Humidity	0%	102%
Specific Humidity	0 g/kg	35 g/kg
Station Pressure	300 hPa	1100 hPa
Wind Speed	0 m/s	50 m/s
Wind Direction	0°	360°
U Wind Component	-50 m/s	50 m/s
V Wind Component	-50 m/s	50 m/s
Precipitation	0 mm	250 mm
Snow Depth	0 cm	300 cm
Incoming Shortwave	-30 W/m^2	1500 W/m^2
Outgoing Shortwave	-30 W/m^2	300 W/m^2
Incoming Longwave	100 W/m^2	500 W/m^2
Outgoing Longwave	170 W/m^2	600 W/m^2
Incoming PAR	-30 W/m^2	2500 W/m^2
Outgoing PAR	-30 W/m^2	200 W/m^2
Net Radiation	-250 W/m^2	1000 W/m^2
Skin Temperature	-90°C	70°C
CO2 Flux	-75 μMoles/m^2/s	75 μMoles/m^2/s
Sensible Heat Flux	-150 W/m^2	600 W/m^2
Latent Heat Flux	-150 W/m^2	600 W/m^2
Soil Heat Flux	-100 W/m^2	300 W/m^2
Soil Temperature	-75°C	75°C
Soil Moisture	0%	100%

Once the format was verified in great detail and the parameters were verified to have reasonable values for somewhere on the surface of the Earth, there remained the possibility of a number of other types of problems. These other types of problems included improperly derived parameters, parameters in incorrect columns, sensor calibration issues, incorrect units, and general data quality problems. In an effort to discover any remaining problems with the submitted data sets, the CDA conducted a thorough visual examination of time series plots of every parameter and flag at every height in every data set from every station from every reference site (e.g. Fig. 6). For a reference site such as the GEWEX Americas Prediction Project (GAPP) Atmospheric Radiation Measurement (ARM) Southern Great Plains (SGP) site that was comprised of several different stations and had observations at several heights, the number of plots examined can reach up to 700. An example of the types of plots examined and problems found can be seen in Fig. 7 and 8. The precipitation data initially submitted to the archive by the reference site is shown in Fig. 7. The maximum value for any 30-min period for the entire 6-month period was 1.5 mm. The CDA questioned the validity of these reported values since this reference site was located in a region that typically has much higher precipitation. The CDA returned the data to the data source for verification. The source determined that they had incorrectly accumulated their high-resolution precipitation data to 30-min values. Fig. 8 shows the precipitation at the site using the correct accumulation methodology. Maximum values now ranged up to 40 mm. This discrepancy was not determined by the gross limit checks since the initially provided values were reasonable for a location on the Earth.

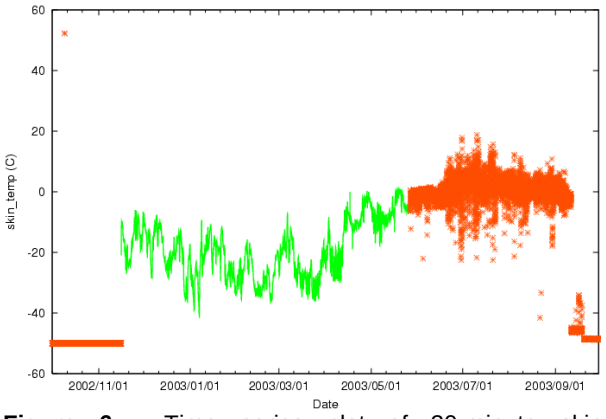

Figure 6. Time series plot of 30-minute skin temperature data at one of the reference site locations. The colors indicate the data quality flag applied for that observation (green is good, orange is dubious, and red is bad).

Additional manual checks are conducted to ensure that the documentation file is consistent with the data file including the reported station location, the reported reference site and station names, and the reported heights of sensors. The results of the automated checking software are examined in detail. The CDA examines the usage of the data quality flags and ensures that they have been properly applied and that significant data issues have been described in the documentation files. Through this suite of detailed and thorough automated and manual checks, the CDA was able to significantly enhance the consistency and quality of the annual cycle data sets and associated documentation.

Following the application of these automated and manual checks by the CDA any potential problems or discrepancies that have been discovered are related back to the data provider for verification and, if needed, correction of the problems. Once a data set has been submitted to the CDA, the entire CDA quality assurance process typically takes days to a week. Once both the CDA and reference site agree the data are of the highest possible quality, are in the proper format and are sufficiently documented, they are released to the community as discussed in the next section.

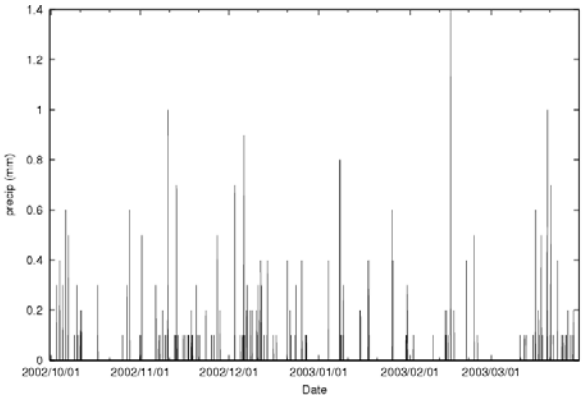

Figure 7. Time series plot of 30-minute precipitation (mm) reported at one of the reference site locations when first submitted to the CDA for the period from 1 October 2002 to 31 March 2003.

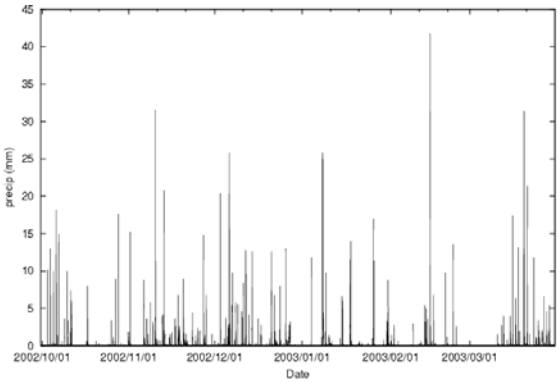

Figure 8. Time series plot of 30-minute precipitation (mm) reported from the same reference site location in Figure 7 after being resubmitted following correction of the accumulation methodology.

5. DATA SET AVAILABILITY

Usage of CEOP data sets are subject to the CEOP Data Policy (CEOP 2003) that, in short, follows WMO Resolutions 40 and 25, prohibits the commercial exploitation of the CEOP data sets and limits transfers to third parties. Additionally, when data from specific reference sites are utilized there are acknowledgement and citation requirements.

The CEOP EOP-1 data sets in both their raw format and in the converted EOP-1 formats are completed and available from the CDA.

As of 27 October 2004 the CEOP first half EOP-3 (1 October 2002 to 31 March 2003) data sets are available from 19 of the reference sites. The CEOP second half EOP-3 (1 April to 30 September 2003) data sets are available from 6 of the reference sites. Additionally, raw format high-resolution (varying from 2-sec to 10-sec vertical resolution) radiosonde data are available from 11 of the reference sites. The CEOP EOP-4 data sets will start becoming available later in 2005.

All of the in-situ reference site data sets, complete documentation and additional supporting information are available on-line from the CDA at: www.joss.ucar.edu/ghp/ceopdm/.

6. ACKNOWLEGEMENTS

The work of the CDA would not be possible without the conscientious efforts of each of the reference site managers and instrument technicians to provide their high quality data sets for the CEOP community. Additionally, without the software engineering expertise provided by Don Stott and Janine Goldstein both of UCAR/JOSS, the timely completion of the proof-of-concept EOP-1 data sets would not have been possible.

7. REFERENCES

CEOP, 2001: Coordinated Enhanced Observing Period (CEOP) Implementation Plan. IGPO Publication Series No. 36, 113 pp. [Available online at www.gewex.org/ceop/ceop_ip.pdf]

CEOP, 2003: CEOP Reference Sites Data Release Guidelines. [Available online at: www.joss.ucar.edu/ghp/ceopdm/ceop_policy.html]

CEOP, 2004: Reference Site Data Set Procedures Report. [Available online at www.joss.ucar.edu/ghp/ceopdm/refdatareport/]

Koike, T., 2004: The Coordinated Enhanced Observing Period – an initial step for integrated global water cycle observation. WMO Bulletin, **53,** 115-121.

Lidard, C. P., and Coauthors, 2004: Comparison of Land Information System (LIS) output with data from the Coordinated Enhanced Observing Period (CEOP). *Eos Trans. AGU,* ***85*** (17), Jt. Assem. Suppl., Abstract H53C-06.

Lu, C, and K. Mitchell, 2004: Land Surface Processes Simulated from the Noah LSM in the NCEP Global Model and North American Land Data Assimilation System: A Comparative Study using the CEOP EOP-1 Reference Site Observations. *Eos Trans. AGU,* ***85*** (17), Jt. Assem. Suppl., Abstract H53C-04.