CHAPTER 2

DATA AND METHODS

"Errors using inadequate data are much less than those using no data at all."

— Charles Babbage, circa 1850

2.1 Datasets

2.1.1 OLR The primary data used in this study are the outgoing long-wave radiation (OLR) from the National Oceanic and Atmospheric Administration (NOAA) polar-orbiting satellites (Gruber and Krueger 1984). OLR is regularly used to locate areas of deep tropical convection and as a proxy for precipitation (Arkin and Ardanuy 1989) by identifying the presence of cold cloud tops. Although there is a good correlation between low OLR values and deep convection in the tropics, it is not possible to distinguish cold cloud tops due to convection from those due to thick cirrostratus decks. This limitation is particularly important in a case-study application and less important when OLR is time-averaged. The OLR data is available from June 1974 through the end of 1997, except for most of 1978 when the data is missing due to satellite problems. The data are interpolated in space to remove any

missing values excluding the extended 1978 gap (Liebmann and Smith 1996). The raw OLR data is mapped onto a 2.5° x 2.5° global grid and represents the average of twice daily (one daytime and one nighttime) satellite passes. Anomaly timeseries are generated by removing the mean and the first three harmonics of the climatological annual cycle (365.25, 182.625, and 121.75 days). It is useful to note that, at least in the Eastern Hemisphere tropics, a 10 W m^{-2} OLR anomaly is approximately equal to a 2.0 mm day^{-1} rainfall anomaly (Xie and Arkin 1998).

A fundamental limitation of the OLR dataset is that the daily average value is obtained from only two satellite passes that occur at the same time each day and night and therefore does not fully resolve the diurnal cycle. In regions where the diurnal cycle in convection is strong, particularly over land (Meisner and Arkin 1987), this limitation will lead to systematic errors in the OLR estimate. The diurnal cycle of convection can only be properly addressed by frequent sampling around the clock with the use of geostationary satellites. Although the monsoon region has been monitored by an Indian government geostationary satellite, the data are not yet available to the scientific community. Recently, the geostationary European Meteosat-5 satellite was moved over the monsoon region which will allow a detailed examination of the diurnal cycle of convection and will hopefully permit the determination of suitable correction factors for daily average OLR estimates.

2.1.2 NCEP/NCAR Reanalysis The National Center for Environmental Prediction and National Center for Atmospheric Research (NCEP/NCAR) reanalyses (Kalnay *et al.* 1996) are employed to represent the dynamically-consistent circulation at 200-mb, 850-mb, and 1000-mb. The reanalysis uses a "frozen" data assimilation model for the entire 1958 to 1997 period which eliminates spurious discontinuities that arise from changes in the assimilation. The reanalysis

data are mapped onto the same 2.5° x 2.5° grid as the OLR data and are available at four times daily resolution from 1958 to 1997. Here, only the 1974 to 1997 subset of daily averaged data is used. It is important to note that OLR and the NCEP/NCAR reanalyses are essentially independent datasets since OLR information is not used to infer diabatic heating or divergent flow.

2.1.3 SST The sea surface temperature (SST) is represented by the Reynolds sea surface temperature analyses (Reynolds and Smith 1994) which are an optimally-interpolated blend of satellite, buoy, and ship measurements. Weekly SST analyses are available from 1982–1997 on a 1° x 1° global grid. Prior to 1982, monthly SST estimates are derived from reconstructed monthly mean SST (Smith *et al.* 1996) that are based on in situ observations and are interpolated with EOFs.

The state of ENSO is evaluated here using the Niño3 SST index, which is calculated by averaging monthly Reynolds SST over the domain 5°S–5°N, 150°–90°W. Numerous other SST indices including Niño 1+2 (eastern Pacific) and Niño 3+4 (western Pacific) are sometimes used to represent ENSO. The Niño3 index is considered typically to be the most representative of the large-scale behavior of ENSO.

2.1.4 Precipitation A daily blended station data/MSU precipitation rate dataset (Magaña, private communication) is used as an independent representation of rainfall. In this dataset, precipitation over land is determined from the historical record of precipitation gauge measurements while precipitation over the oceans is inferred from the Microwave Sounding Unit (MSU) data (Spencer 1993). The blended precipitation dataset is available globally on a 1° x 1° grid from 60°S–60°N for the period 1979 to 1995.

The seasonally averaged strength of the South Asian monsoon is assessed with the all-India rainfall index (AIRI) which is available from 1871–1997. The AIRI is a weighted average of 306 well distributed rain-gauge stations across India (Parthasarathy *et al.* 1992; Parthasarathy *et al.* 1994). Further details on the AIRI and its relation to regional rainfall anomalies can be found in Mooley and Parthasarathy (1984). Descriptions of other measures of South Asian monsoon strength and a brief analysis of their interannual variability are found in Section 4.3.1.

2.2 Methods

2.2.1 Statistical Techniques
The primary conclusions drawn in this study are based fundamentally on statistical analyses of South Asian monsoon variability. As such, an extensive suite of statistical techniques are employed which facilitate exploration of the interrelationships between ISO convection and circulation, interannual variations of intraseasonal convection and the seasonal strength of the monsoon, and the synoptic-scale convection and the ISO. The statistical techniques used include the following:

- lagged cross-correlation and linear regression,
- Fourier spectral analysis and temporal filtering,
- wavenumber-frequency spectral analysis and filtering,
- wavelet analysis and wavelet derived variance,
- empirical orthogonal function (EOF) analysis, and
- composite analysis.

The details of each statistical technique are incorporated within the body of the thesis at the point of first use, except for the cross-correlation and linear regression method which is described in the following section.

2.2.2 Lagged Cross-Correlation and Linear Regression The cross-correlation and linear regression technique used here is the same as that described in detail by Kiladis and Weickmann (1992). Area-averaged OLR is used to represent tropical convection in a base region. The OLR timeseries is typically extracted from a bandpass filtered OLR dataset (the bandpass filtered datasets used here and motivations for their selection will be discussed in Sections 3.2 and 5.2) so that particular spatial and temporal scales of convection can be isolated. The base region bandpass filtered OLR timeseries is regressed against identically filtered OLR and 1000-mb, 850-mb, and 200-mb u (zonal wind component) and v (meridional wind component) timeseries at all other grid points. Since the focus of this study is on boreal summer, regression equations are primarily developed based on convection in the June–September period. However, for the purpose of comparison between the summertime and wintertime ISO, regression equations are also developed for the northern winter months, December–March (Section 3.3). The linear dependence of the circulation on deep tropical convection can then be mapped by applying the regression equation at each grid point relative to a single standard deviation in OLR. The temporal evolution of the circulation and convection anomalies can be assessed subsequently by performing lagged regressions.

This technique assumes that the relationship between OLR and the circulation is nearly linear, an assumption that is reasonable to the extent that OLR anomalies are related linearly to tropical heating anomalies and that linear dynamics can describe the response to tropical heating. A number of studies have suggested that, in the tropics, realistic circulation perturbations can be generated by forcing linear barotropic models with steady divergence sources (e.g., Matsuno 1966; Webster 1972; Gill 1980; Sardeshmukh and Hoskins 1988). However, a caveat

to the assumption of a predominantly linear response to tropical heating is that non-linear interactions, such as the feedback of the circulation on the convection itself, is likely to be a significant factor in the total response to convective heating (Chang and Lim 1988).

The statistical significance of the local linear relationship between OLR in the base region and the dependent variable at each grid point can be determined using the correlation coefficient. For all linear regression maps shown in this study, if the absolute value of the correlation coefficient between the OLR base region timeseries and the dependent variable timeseries at a particular grid point exceeds the 5% significance level correlation coefficient, then the regressed value for a negative one-standard-deviation (σ) base region OLR value is plotted at that grid point. Similarly, if the absolute value of the correlation between the OLR base region and either the u- or the v- wind component at a single grid point exceeds the 95% significance level correlation coefficient, the total wind anomaly vector is plotted at that grid point.

A sample cross-correlation and linear regression map is shown in Fig. 2.1. Scatter diagrams of 25 to 80 day filtered u_{850}, v_{850}, and OLR at 65°E, 10°N in the Arabian Sea versus OLR in a base region located in the central Bay of Bengal (85°–90°E, 12.5°–17.5°N) are shown in Fig. 2.1a for lag zero days. Figure 2.1b shows the global map of regressed 850-mb wind vectors and OLR at lag zero days. The correlation coefficients for the zonal and meridional wind vectors at the Arabian Sea grid point are –0.55 and –0.44 respectively. A correlation coefficient whose magnitude exceeds 0.25 is statistically significant at the 95% level. Therefore, since the correlation coefficients for both u_{850} and v_{850} are greater than 0.25, the total wind vector is plotted in Fig. 2.1b. The correlation coefficient for OLR is only +0.13 and
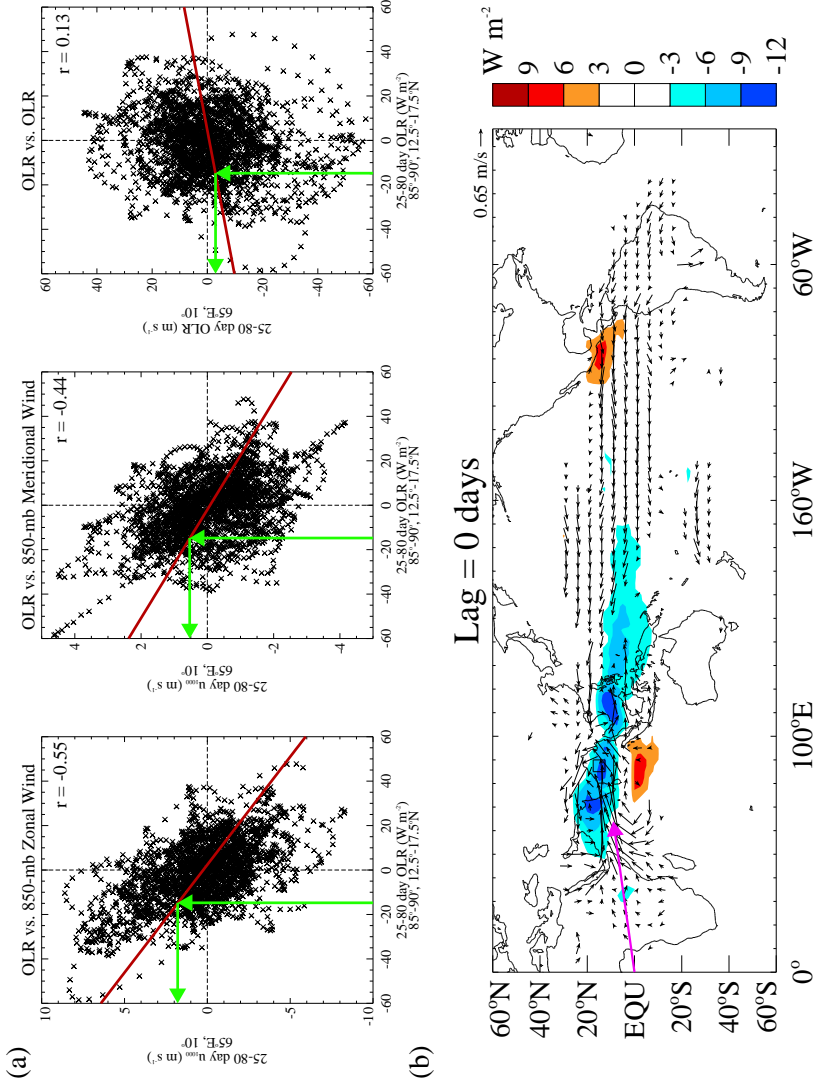
Figure 2.1. Example of cross-correlation and linear regression for summertime OLR bandpassed to retain periods 25 to 80 days. Base region is 85°–90°E, 12.5°–17.5°N. (a) Scatter diagrams of 25–80-day filtered OLR in base region versus 25–80-day filtered u_{850}, v_{850}, and OLR at 65°E, 10°N. Vertical green arrow points from -1σ base region OLR value to fitted line. Horizontal green arrow points from intersection of vertical green arrow and fitted line to regressed grid point value. (b) Map of regressed 850-mb wind (vectors) and OLR (filled contours) for a -1σ in base region OLR at lag 0 days. Only wind vectors and OLR that are locally significant at the 95% level (correlation coefficient greater than approximately 0.25) are plotted. The OLR base region is denoted by a black box. Violet arrow points to grid point detailed in (a).

therefore is not statistically significant. Consequently, the regressed OLR value is not plotted at 65°E, 10°N.