Description of the Predictor Data Sets used in RegClim's Statistical Downscaling Project

R.E. Benestad

DNMI, Oslo, September 4, 1998

Contents

1 Introduction

A number of data sets will be used for statistical downscaling in the RegClim[1] project, whose objective is to predict likely regional climate changes in Scandinavia as a result of a global warming scenario. The purpose of this report is to give an overview and evaluation of the different data sets used as predictors in the statistical downscaling project. The differences between the data sets, of for instance SLP, can give an indication about the uncertainties in the gridded data fields, and hence provide a basis for estimating the reliability of the statistical model approach. Only data sets of monthly mean values will be discussed here.

The evaluation of the gridded data will involve statistical methods, such as Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA), applied to corresponding data fields from different data sets. The spatial patterns of variability from the PCA will be henceforth referred to as Empirical Orthogonal Functions (EOFs). The PCA may be regarded as an extension of an ordinary regression analysis, $\vec{y} = \vec{x}A$, to data fields with many dimensions, and CCA may be thought of as a multivariate extension of correlation analysis. Thus the PCA aims to minimize the RMS errors[2], or equivalently find spatially coherent patterns of variability that maximize the variance under the constrain that each EOF in a given set is orthogonal to all other EOFs in the same set. The CCA finds the spatial patterns in two data sets that have the greatest correlation in time.

The EOF (or CCA) patterns are expected to be similar if all the data sets only contain small errors. Differences between these patterns may indicate that one or more of the data sets have errors in them, but can also be a result of different grid resolution. Such information can provide a crude measure of the errors in the predictors and the uncertainty in the downscaling products.

A brief overview of the different data sets used in RegClim's statistical downscaling will be given first, followed by a more detailed description of each data set. A comparison between the gridded and the station (observations) data will be discussed in the following section.

The data sets discussed here are available at DNMI as well as on the world wide web (see RegClim's intranet page: http://gust/regclim/).

[1] Regional Climate Development under Global Warming

[2] Here, "errors" refer to departure from the "true" value due to inaccurate observations or low resolution, but "bad" values as a result of bugs in the analysis or mistyping will also be included in this definition.

Table 1: *Overview of the gridded Predictor Data sets. The following legends have been used: "SLP" for Sea Level Pressure, "SST" for Sea Surface Temperature, "TA" for Surface air Temperature, "T(z)" for Temperature at a specific pressure level, "Z(z)" for geopotential height, "MM" for Monthly Mean values, and "ID" for Instantaneous daily observation*

Data set name	**Quantities**	bf Resolution
Advance-10k	SLP, TA	$5° \times 5°$ MM
COADS	SLP, SST, TA	$2° \times 2°$ MM
GISST2.2	SST, Sea ice	$1° \times 1°$ MM
NCAR ds010.0	SLP	$5° \times 5°$ ID
NMC (NCEP) ds195.5 CD-ROM	SLP, T, Z	$5° \times 5°$ ID
University of East Anglia (UEA SLP)	TA	$10° \times 5°$ MM

2 An overview of the data sets

2.1 Description of the different data sets.

The different data sets used as predictors in *RegClim*'s statistical downscaling are described in an alphabetic order below. Each sub-section gives a general description about the data, the file format, and information on how to access the data.

2.2 Advance-10K data.

2.2.1 General information

The EU project, *Analysis of Dendrochronological Variability and Natural Climates in Eurasia: THE LAST 10,000 YEARS (Advance-10K),* produced two data sets, containing SLP and surface air temperatures respectively. The original data are archived as ASCII files, but with different formats for the SLPs and temperatures. Local versions of the temperature data set have been archived in a commonly used scientific data format called *netCDF*. The size of the *netCDF* files are less than 3Mb depending on regional or hemispheric version, and the ASCII files take up approximately 22Mb in total.

The temperatures are stored as monthly means on a $5°\times5°$ grid for the time period Jan. 1854- Dec. 1995. The data format of the temperature ASCII files is one block for each month, consisting of a header of two numbers (2i6), followed by a block of 18x144 (18i5)= 2592 data points. The number of spatial points are 72x36, and the spatial coverage is for the northern hemisphere. The units of the data in the original data files are in centi degrees Celsius (1/100 °C).

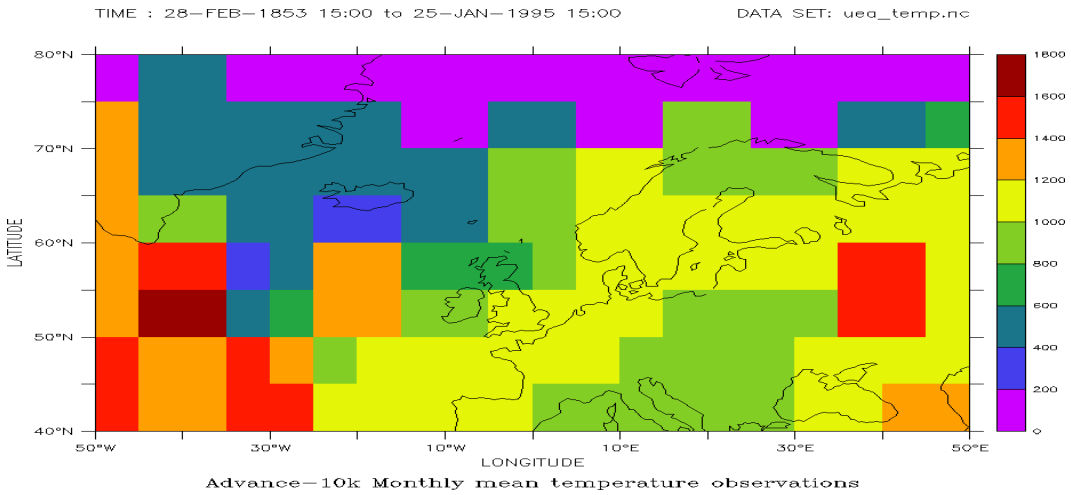

Figure 1: *The spatial distribution of the number of valid observations in the Advance-10K monthly mean temperature data set. The regions with best temporal coverage, i.e. longest time series, are indicated as red shading (such as in the mid latitude Atlantic) and the regions with few observations are shown in blue (such as in the arctic seas) and purple for no valid data over land.*

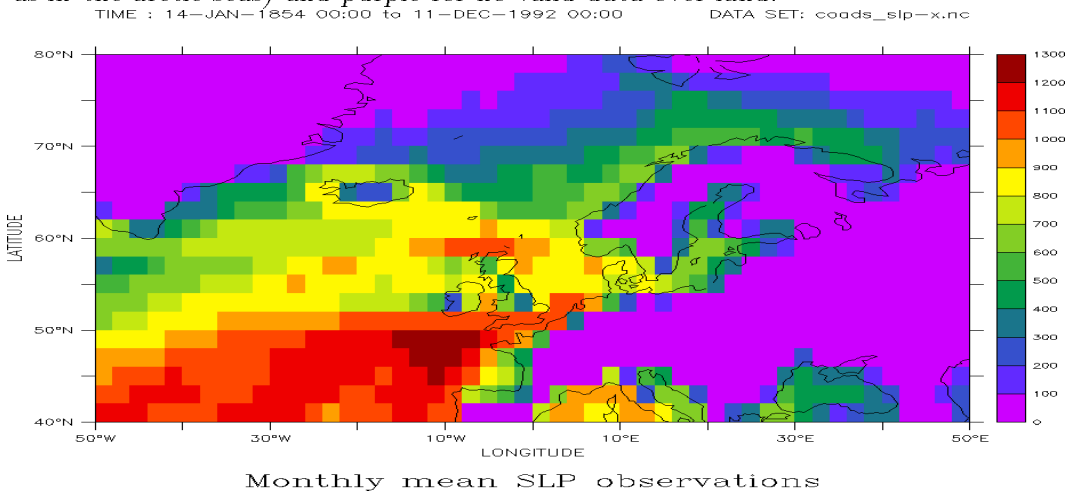

Figure 2: *The spatial distribution of the number of valid observations in the COADS monthly mean SLP data set. The regions with best temporal coverage, i.e. longest time series, are indicated as red shading, such as in the mid latitude Atlantic, and the regions with few observations are shown in blue, such as in the Arctic seas, and purple for no observations over land.*

Description of the file format for the Advance-10K SLP was not available on the internet site, and for this reason, the UEA SLP data (see below) were used instead of the Advance-10K SLP.

Figure 1 shows a map of the number of observations with valid data. The data set has a relatively good coverage over the higher latitudes. More information about the data set can be found at the URL:
http://www.cru.uea.ac.uk/advance10k/ and The *Holocene, vol. 2 no. 2 (1992).*

2.2.2 How to read the data?

The data set will eventually be stored on CD-ROM (*DNMI* library) and on saragasso.oslo.dnmi.no. Any application that can read data in ASCII format, such as *Excel* and *SAS*, in addition to *FORTRAN* and *C++* programs, can be used to read the data. *Excel* is, however, not ideal for reading gridded data, as it doesn't manage the shear volume, and *SAS* does not easily cope with geophysical data fields with more than 2 dimensions. *FORTRAN* and *C++* programs involve a lot of programming and testing which is rather time consuming. The data are also archived in the *netCDF* format, which is in general easier to deal with than most other data formats provided one has the right software. It is not necessary to know much about the format; it is a transportable[3] format of self-describing data files[4] with direct access[5]. There is an abundance of (free!) scientific programs that read and plot *netCDF* for *UNIX/Linux* platforms. One such package is called *Ferret*, developed by *NOAA/PMEL/TMAP* in the US, which is now installed on front and saragasso. *Ferret* is probably the quickest and simplest package to use as it is tailored for geophysical data sets and data visualisation. *GrADS* is another package that can read *netCDF* (only latest version on the *UNIX* platforms). *Matlab* (version 5.0 and later) can also read the files, and a number of routines have been written for this purpose (ncread.m) at DNMI are located on the DNMI *Klimaavdelingen*'s Matlab intranet page: http://gust/regclim/matlab_scripts (In order to run these with *Matlab*, mexcdf and *netCDF* toolbox must be installed, which are freely available from the web. These have also been installed locally on saragasso at DNMI). The

[3]Which means that the data can transferred the between Crays, Workstations, PCs and Macs without problems or need for bit swapping.

[4]The files contain a header with metadata in addition to the data itself, so that the file contains all the necessary information on how to read the data.

[5]Direct access means that it is possible to pick out a subsets without having to read all the preceding data and generally makes it faster to read the data than, say *GRIB* and formatted files.

command UNIX *ncdump* displays the content of the *netCDF* file to screen (usage: *ncdump* { *filename* } | *more* . Here, the text with curly brackets,"{ ..}", should be replaced by the file name). It is also easy to produce ASCII files from the *netCDF* format by using the UNIX command line: *ncdump* -v { *variable* } { *netCDF_filename* } > { *ascii_filename* } . For example:
ncdump -v *TA uea_temp.nc* > *uea_temp.asc*

It is conventional to use the suffix '*.nc' or '*.cdf' for the *netCDF* files, and we will employ the '*.nc' notation here. The ASCII files will use either of the suffices '*.dat', '*.asc'.

2.3 COADS SST, SLP, Surface air temperatures, and surface relative humidity.

2.3.1 General information

The *COADS* (Comprehensive Ocean-Atmosphere Data Set) data are the product of a cooperative project which had as a goal to create consistent and easy-to-use historical record of marine surface data. The members of this U.S. based project included National Climatic Data Center (*NCDC*), Environmental Research Laboratories, the cooperative Institute for Research in Environmental Sciences, and the National Center for Atmospheric Research (*NCAR*).

The data are stored on a global $2° \times 2°$ grid in the *netCDF* format. The *COADS* data contain monthly mean values estimated from marine ship observations ("ships-of-opportunity") and buoy data for the period 1845-1992. The data have been quality controlled, and additional files describing number of observations and standard deviation per grid are available at the *COADS* data site. The limitation of the *COADS* data set is that it only contains maritime data, i.e. no data over land. Furthermore, regions with sparse information and few ship observations may suffer from degraded quality or missing data. There is a relatively poor data coverage in the Arctic seas. It is unlikely that the other data sets have much more observations in the Arctic, although missing data gaps may not be as obvious as in the COADS data. It is common to patch holes in the data set by interpolation or use of models. Figure 2 therefore gives a good idea of how well the higher altitudes actually is sampled in most of the data sets.

A detailed description of the *COADS* data is given in the reference: *Slutz* et al. (1985) and on the URL:
http://ingrid.ldgo.columbia.edu/SOURCES/.COADS/.dataset_documentation.html.

2.3.2 How to read the data?

Description of how to read the data in *netCDF* files is described in the *Advance-10K* section. There are two types of *COADS* files: 1) coads_slp.nc and coads_sst.nc with global coverage. These are the original copies that were downloaded from the internet (they are not saved according to the 'standard' *netCDF* conventions described in the *netCDF* and *Ferret* documentation); 2) The smaller subsets of region 40°N-80°N, which are named coads_slp-x.nc and coads_sst-x.nc (which follow the *netCDF* conventions used by *Ferret*). The data files will in general be stored in a compressed form, indicated by the extension *.gz, and can be decompressed by tools such as *gunzip* (*gzip -d*) on *UNIX* platforms and *WinZip* on *MS Windows* systems.

The original COADS files sizes are 108Mb for the air temperature (TA), relative humidity (RH), sea level pressure (SLP), and the sea surface temperatures (SST) respectively when these files are not compressed. The local versions of these files are approximately 15 Mb large, partly due to an extraction of a smaller region, but also because these files store the main bulk data as short type and use scaling and offsets (given in the file header) for reconstruction of the original value.

2.4 The NMC gridded analysis: ds195.5

2.4.1 General information

The *NMC (NCEP) ds195.5* data set comes on 2 CD-ROMs, and contains twice daily data values (00 and 12 UTC), climatological, and monthly mean data. The gridded data are based on the NMC final analysis and include data received up to about 10 hours after data time. In some cases, when the final analysis was not available, the data was filled in with operational data. The generation of the final data was based on several different methods which include: Cressmann objective analysis, Hough Analysis, Global Optimum Interpolation Analysis, Global 24-mode Spectral Model with 12 layers, Global 30-mode Spectral Model with 12 layers and T126L18 Spectral Forecast Model. In other words, some of the gridded values are generated using dynamical atmospheric models that use observations as initial and boundary values.

The CD-ROMs contain Sea Level Pressure (SLP), Geopotential height (Z), Temperature (T), horizontal wind components (U, V), vertical wind component (W), and relative humidity (RH) at following levels: Sea-level, 850hPa, 700hPa, 500hPa, 250hPa and 100hPa (not every field is given at all levels). The relative humidity is given for the 1000-666 hPa layer. The spatial coverage is the Northern Hemisphere, and the grid is octogonal (47x51), with a spatial resolution that is greater than 5° for latitudes higher than 40N. The longitude-latitude grid has a resolution of 2.5°× 2.5°, which is higher than for the other SLP data sets. The data set covers the period 1946-1994, but can be updated to 1996 (via ftp). Some data fields [i.e. Z(850hPa) and Z(700hPa); R.H.] have no valid data before

1962 and 1973. Virtual temperatures were apparently archived instead of real temperatures after September 1978, but it is unclear what happened to the analysis model at this time. There appears to have been some model changes between 1974 and 1978, and it is not entirely clear exactly when virtual and when real temperatures were archived. Data of uncertain quality have been flagged where the values have been quality controlled and the data values are questionable. The most significant discontinuity reported involves a change of interpolation scheme from 00:00Z-18-Nov-1986, which may have affected the 700hPa geopotential height field and 700hPa temperatures near the Himalayas.

2.4.2 How to read the data?

The ds195.5 CD-ROMs are available at the DNMI, and includes routines that convert the orthogonal grid to regular longitude-latitude grid. These routines do not run very well on the Irix6.3 platform, however, a FORTRAN program was obtained from NCAR (by Chi Fan: chifan.f) which converted the packed NCAR on an octogonal grid to regular longitude-latitude daily grids in binary files (approximate size is 400Mb). A second FORTRAN program (In the bin2ncdf.sh shell, which both compiles and runs the code) was then used to convert these binary files to *netCDF* files of monthly mean (5Mb) and daily values (130Mb).

Some documentation for the data set was given in a text file on the CD-ROM and a User's manual *NMC* (1996), however, there were no other references to any publications based on these data.

2.5 The NCAR gridded analysis: ds010.0

2.5.1 General information

The *NCAR ds010.0* data set contains the *NCAR* daily SLP values for the period 1899-1997. The data are stored on a 63x63 octogonal grid that covers the Northern Hemisphere and extends as far south as 20°N. The data is also available on a 72x15 longitude-latitude grid with 5°×5° resolution. The *NCAR ds010.0* SLP data set contains daily instantaneous values that correspond to 13:00hZ between 1899 and 1939, and instantaneous values of SLP corresponding to 12:00hZ between 1939 and 1956 (see table 2). After 1956, the values are stored for both 00:00hZ and 12:00hZ. The *netCDF* files, however, only contain the 12:00hZ and 13:00hZ fields.

The sources of the data are ship observations, routine operational observation or analysis, and ESSPO[6]. Although the *NCAR ds010.0* data has been quality controlled, there may still be some inhomogeneities due to the fact that the observation time has changed from 13:00hZ to 12:00hZ (table 2) and that the data originates from different sources (using different observational platforms/analysis)

[6]Apr 1955 - Mar 1960: Points were hand read from historical charts by the 433L ESSPO Project. The sources of the ds010.0 data set are from MIT, NOAA, NMC or NCEP, and the US Navy

Table 2: *Overview of the NCAR ds010.0 data set sources.*

Date	Sources	Remarks
Jan 1899 - Jun 1939	Historical maps: National Climate Center	13:00Z
Jul 1939 - Nov 1944	MIT: Extended forecast lab	12:00Z
Dec 1944 - Dec 1945	(NONE due to World War 2)	
Jan 1946 - Mar 1955	U.S. Navy: Historical series maps (occasionally NMC grids)	12:00Z include manual: tropical storms
Apr 1955 - Mar 1960	ESSPO	00:00,12:00Z
Apr 1960 - Jun 1962	U.S. Navy: Digitized with a curve follower (NCC) (occasionally NMC grids)	00:00,12:00Z include manual: tropical storms
Jul 1962 - Feb 1998	U.S. Navy: Navy's operational objective analysis (occasionally NMC grids)	00:00,12:00Z include manual: tropical storms

in different periods. The grids from the U.S. Navy included manual bogus to put in tropical storms. An error was made (at *NCAR*) in the construction of the gridded data sets where 235°E -355°E were duplicates of 0°E - 120°E after 1994, however, the later versions of the data have this corrected (a corrected data set for 1994 to end of Feb 1998 was provided).

2.5.2 How to read the data?

The FORTRAN code to read and write the ds010.0 data as *GrADS* and ASCII files was obtained from the internet, and these programs ran on the SGI Irix6.3 system after a few alterations. (The FORTRAN code is written in the 1970's spirit which is difficult to follow due to the use of "spagethi programming style", a large number of nested "go to"s). The *netCDF* files (converted using the code ds2ncdf.sh) can be read using *Ferret* or *Matlab*, and the monthly mean version take about 5Mb.

2.6 UEA MSLP

2.6.1 General information

The monthly mean SLPs in the *UEA MSLP* data set are based on *Jones* (1992), and cover the period Jan. 1873-Dec. 1995. The data grid has a spatial resolution of 10°×5° (lon-lat) and covers the area from 0°E-10°W to 15°N-85°N. The original data is archived as ASCII files, however, a conversion has been made to *netCDF* at

DNMI (read_uea_slp.m). The original data were stored in blocks of monthly mean values in the following layout: a header of 5 numbers (5i6) and 576 data points (36x16). The values were stored in the original ASCII files as x, where SLP (hPa)= $x/100 + 1000$, and missing values were set to -32768. The *netCDF* versions have the missing value, the offset and scaling information in the file header (as variable attributes). References on these data sets are given by: *Basnett & Parker* (1997), *Trenberth & Paolino* (1980), and *Jones* (1992). An URL with more information can also be found on:
http://www.cru.uea.ac.uk/cru/data/pressure.htm.

Figure 3 shows the spatial data coverage in number of observations. The data set contained 122 years of data, which below 70°N gives 1464 observations.

2.6.2 How to read the data?

The data can be found on saragasso.oslo.dnmi.no in the *netCDF* format which takes 1.6Mb. See the *Advance-10K* section for description of this format. There are some *Matlab* routines that also read the original ASCII files (The *Matlab* routines can be found on the intranet address: http://gust/regclim/matlab_scripts/). The *Matlab* routine that reads the ASCII file is called read_uea_slp.m.

2.7 UKMO GISST2.2

2.7.1 General information

The United Kingdom Meteorological Office (*UKMO*) global ice and sea surface temperatures (*GISST*) 2.2 data set is a derivation from the Meteorological Office historical sea surface temperature archive 6 (*MOHSST6*[7]) and Walsh sea ice data[8]. The monthly *GISST2.2* SSTs are reconstructed from: 1) 1949-1981 monthly mean values *MOHSST6* sea surface temperature anomaly (SSTA) eigenvectors on a 2°×2° resolution; 2) 1982-1994 a blend of *in situ* and satellite measurements on 2°×2° resolution. The reconstruction used the *GISST2.0* climatology (here defined as climatological mean values for the 1961-1990 reference period) with a spatial resolution of 1°×1° as reference: the SSTAs (deviations from this climatology) were subject to interpolation and smoothing and then added to the climatological SSTs.

Figure 4 shows the number of valid data points in the GISST2.2 data set (January month). The GISST2.2 data set are stored on a 1°×1° global grid (size 360x180) of monthly mean values and covers the period January 1903 to December 1994. The original data format is the *Hierarchical Data Format - Scientific Data Set* (*HDF-SD*), but the data on saragasso have been converted to *netCDF*. The SSTs in the *HDF-SD* files are set to -1000 where more than 95% of the area is

[7]The MOHSST6 data set contains some COADS (analysed) data in data sparse regions.
[8]Sea-ice concentration data from a new analysis by J.Walsh, *Rayner* et al. (1996)

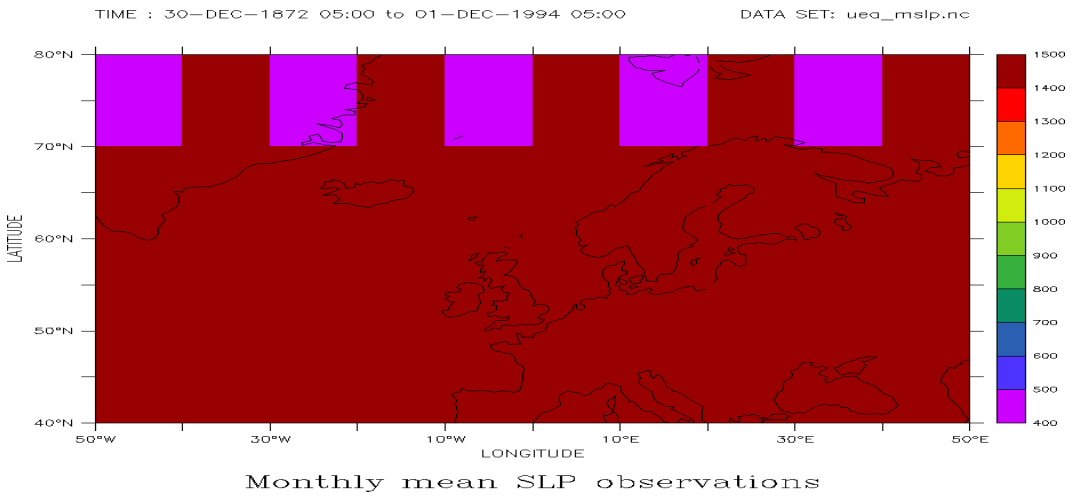

Figure 3: *The spatial distribution of the number of valid observations in the UEA MSLP monthly mean SLP data set. The regions with best temporal coverage, i.e. longest time series, are indicated as brown shading. This data set contains only a few missing data gaps.*

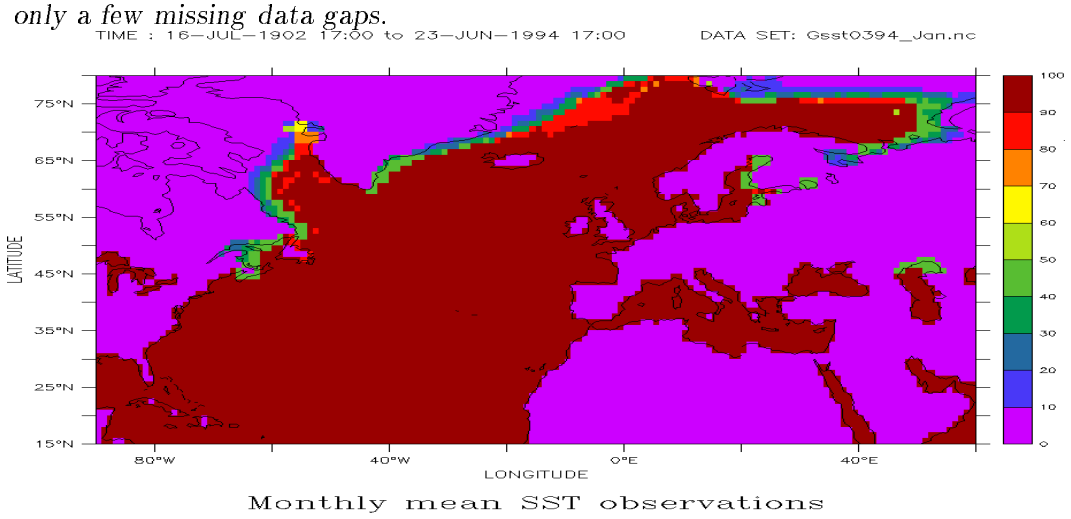

Figure 4: *The same as for figure 1, but for GISST2.2 January SSTs. Notice, the shading values here gives the number of January observations, i.e. number of years of observations.*

covered by ice and land points are set to -32768. The sea ice data (stored in separate files: GiceY1Y2.nc) contain an ice index of 0 to 10, where 0 denotes no ice and 10 means permanent ice cover.

The *netCDF* versions of the GISST2.2 data set contain 20-year sequences of the entire data (4.7Mb), but some files also contain only one calendar month (for instance only January SSTs) for the entire period (1.8Mb). The reason for splitting the data up into several files is because of the memory limitations when analysing the data with *Ferret* or *Matlab*.

More information as well as references are given on the CD-ROM itself. Further details on the data set can be found in: *Rayner* et al. (1996) and *Parker* et al. (1995).

2.7.2 How to read the data?

The data set can be found on the Global Ocean Surface Temperature Atlas Plus (GOSTAplus) CD-ROM, which is available for the *RegClim* project at *DNMI*. Software, such as HDF-to-ASCII converters and visualisation tools are also available on the CD-ROM. *Matlab* (version 5.2 or higher) also can read HDF-SD formatted files directly (using the "hdfsd" command). A *Matlab* code, *hdf2ncdf.m*, converts the *HDF-SD* data to *netCDF*.

3 Data preprocessing

3.1 Anomalies and Climatology

The term *climatology* often refers to the mean seasonal or monthly value of some particular quantity. In this report, climatology is referred to as the monthly value, i.e. the annual mean plus the contribution from the annual cycle (which here is defined as a sinusoid with zero mean that describes the annual variations)[9]. The climatology is then easily estimated by taking the average of the values that correspond to the particular calendar month, i.e. the temperatures for all January months. The current standard period for which the climatology is estimated is 1961-1990, and these climatological values are often referred to as the *standard normal values*[10]. Some data sets may not span the period 1961-1990 (for instance model integration), and therefore the climatology does not always refer to this period.

The *anomalies* are computed by subtracting the climatology from the observed data values. The anomalies do not describe the seasonal variations, but only

[9]This use of "climatology" is commonly used in the dynamical climate modelling community.

[10]We will use the term "climatology" here, and not "normals", because "normal" also has meanings such as "orthogonal" or "unit length" (as in normalised) in terms of vectors (EOFs), or "Gaussian" in terms of normal distribution.

interannual, decadal, and variability with longer time scales.

3.2 Methods for comparing and assessing the predictor data sets

The data evaluation carried out here will mainly focus on the January mean values in order to limit the length of this report. The different methods used for examining and comparing the various the data sets can be summarised as follows:

- The data can be presented as **time series**, where the quantity plotted may be an area average, a measurement from a location or an interpolation in the gridded data set that corresponds to a particular location. Comparison between the time series estimated from the different data sets can indicate differences in terms of mean value, trend, and variance. Differences between the data sets indicate errors in the gridded data, which can be related to particular events. The point measurements may not always represent the most appropriate (interpolation) values for the gridded area means, as regional topography may give rise to biases that are not representative for a larger region, and therefore the differences between these may not necessarily indicate errors in the gridded data when comparing the station data and the gridded data.
- **Empirical Orthogonal Function (EOF)** are the spatial maps from the PCA, where the data have been weighted appropriately by geometric spatial weights. The EOFs can be used to reduce the size of the data if only a small fraction of the EOFs represent the signal and the rest is noise (i.e. by reducing the degrees of freedom). Another useful property of the EOFs is that the different EOFs are orthonormal (orthogonal with unit length), which makes the linear algebra simple. A comparison between the EOFs and corresponding eigenvalues of different data sets also gives information about the strength and location of spatially coherent patterns of variability.
- **Canonical Correlation Analysis (CCA)** yields two sets of spatial maps, one for each data set, that have optimal correlation; The leading CCA pattern pair corresponds to the spatial patterns that have the highest possible correlation for the two data sets. The CCA patterns indicate where one quantity is correlated with another quantity, even though the locations of high correlation may not necessarily coincide spatially.

3.3 Statistically independent realisations and spatial weighting

North et al. (1982) and *Wilks* (1995) state that the time series must consist of statistically independent realisations in time before most statistical analyses can

be carried out. Most of the analysis methods described here assume statistical independence between the different realisations. We apply preprocessing steps to eliminate/reduce the temporal autocorrelation by sub-sampling the data fields. The sub-sampling was implemented by extracting only one of the calendar months and then applying principal component analysis (PCA) to these monthly values. This way, 12 different sets of Empirical Orthogonal Functions (EOFs) were computed to describe the 12 calendar months.

North et al (1982) argued that it is necessary to use sufficiently high spatial resolutions to capture important small scale variability. The spatial resolution of the data sets is $10°\times5°$ for the UEA MSLP data, and the PCA on this data set will not capture small scale variability such as storm tracks with sufficient realism. It is also likely that the data from regions with sparse observational network, such as northern Norway and the Arctic, may contain more (random) sampling errors than those from areas with dense observations such as the UK. On the other hand, the statistical methods may, unless the individual data points are appropriately weighted, give too much weight to variability over more densely station-populated areas and therefore miss out important features in less well covered areas further to the north. It is therefore important to ensure good data quality while simultaneously capturing important processes in the Arctic. One solution to this dilemma is to use an estimate of the error covariance matrix for the observations/gridded data and a geometric scaling factor to estimate the spatial weighting function to be applied to the data prior to the analysis. Further spatial weights (cosine shape) were in some cases applied to emphasis the local area near Scandinavia and suppress the variability in the remote regions.

3.4 Missing data

Data sets in general contain some missing data gaps. In most cases, small missing data gaps have already been filled in for the data sets by some interpolation scheme. The COADS data was extremely sparse in the early record, however, the spatial coverage improved with time. The locations which did not have valid data for all times during the period selected for PCA or CCA analysis were not included in the analysis.

3.5 Comparison between observations from stations and the gridded data sets.

Time series from grid points in the gridded data sets were compared with the observations from stations along the Norwegian coast. The time series from the stations, UEA SLP, NCAR SLP, COADS SLP, and NMC SLP for the location that corresponds to that of *Oksøy fyr* are shown in figures 5, 6 and 7. Although there are some differences, the various data sets show similar features. The NCAR ds010 and NMC ds195.5 values are systematically lower than the observations by

1-2 hPa before 1970, after which the systematic error is reduced significantly (not shown). The UEA SLP values tend to be too high by about 1 hPa. The comparison between the station data and the gridded data suggests that in general the pressure level in the NCAR SLPs is more accurate than UEA after 1970. The SLP time series in the NMC ds195.5 data set have slightly greater amplitude than the the corresponding data in the other data sets, and hence, the NMC data accounts for a larger amount of variance (figure 7).

The NCAR (ds010.0) and the COADS data sets both contain data from common sources, such as ship observations (the former US Navy), and it is not clear whether these two data sets contain independent observations.

SLP time series estimated from the different gridded data sets were compared with station data from the Arctic: Bjørnøya (Bear island) at 74.31°N - 19.01°E (figure 8) and from Northern Norway: Bodø at 67.2°N - 14.3°E (not shown). This comparison shows a good agreement between the station data and the gridded analysis before June 1939. In fact, the time series from the different data sets appear to correspond more closely here than at Oksøy fyr, which lies in a region with more dense network of observations. The explanation of this apparent "paradox" may be that the different data sets have few observations in the Arctic regions, and may include the same station data against which they are compared. The SLPs in regions with no observations have been deduced from analysis, and the analysis tend to rely strongly on the observed values in the vicinity of the station in data sparse regions. Therefore the analysis is in close agreement at the locations where there are observations. The quality of the data where there are no observations is more uncertain. Figure 9 shows the SLP comparison at 75°N - 5°E and illustrates how the different data sets start to diverge away from the observations.

Figure 8 indicates a change in the NCAR SLP characteristics near Bjørnøya during 1939. Before 1939, the NCAR SLPs show low variance compared to the observations from this region, but the variance of the NCAR data increases to similar values as the observations during the summer of 1939. Table 2 shows that the data set switched from using historical maps from the National Data Center to MIT extended forecasts during June 1939. Figure 8 illustrates the problem with data quality in the high latitudes, and this problem is not only limited to the NCAR data set. The UEA data set does not contain valid data near 74°N - 19°E before 1950, and for this reason, the data at the region north of 70°N has been excluded from the PCA described below.

The pressure differences can give an indication of the atmospheric flow as the geostrophic flow is balanced by the pressure gradient. The difference between SLPs at two locations is small compared to the observed pressure values, and it is important to keep in mind the dangers of analysing the difference between two large noisy numbers. The SLP scatter plots (figure 6 and 7) indicate that the errors in the monthly mean SLP values may be as high as 8 hPa (Oksøy fyr) and typical scatter of about 4 hPa, which is of same magnitude as typical SLP differences between Oksøy fyr and Bergen (not shown). For this reason, the

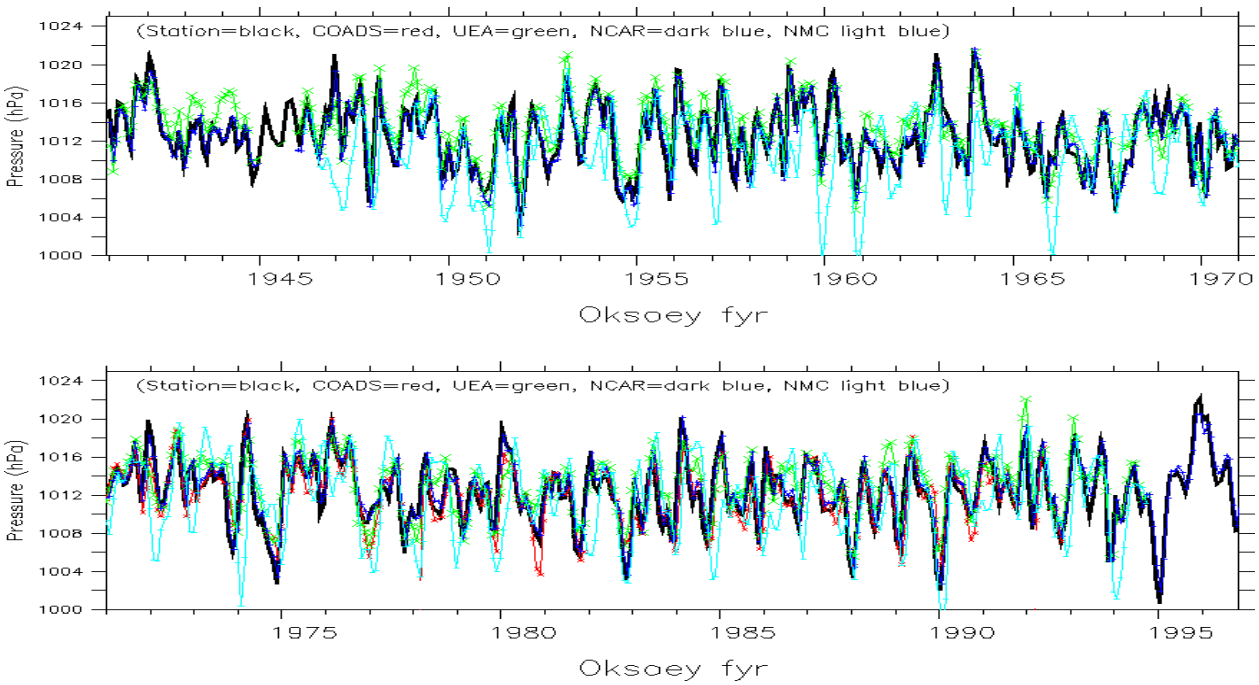

Figure 5: *Time series of monthly mean SLP values from Oksøy fyr and from the predictor data sets (Station observation in heavy black and COADS SLP in red; UEA SLP in green; NCAR (ds010.0) in dark blue; NMC (NCEP: ds195.5) in light blue).*

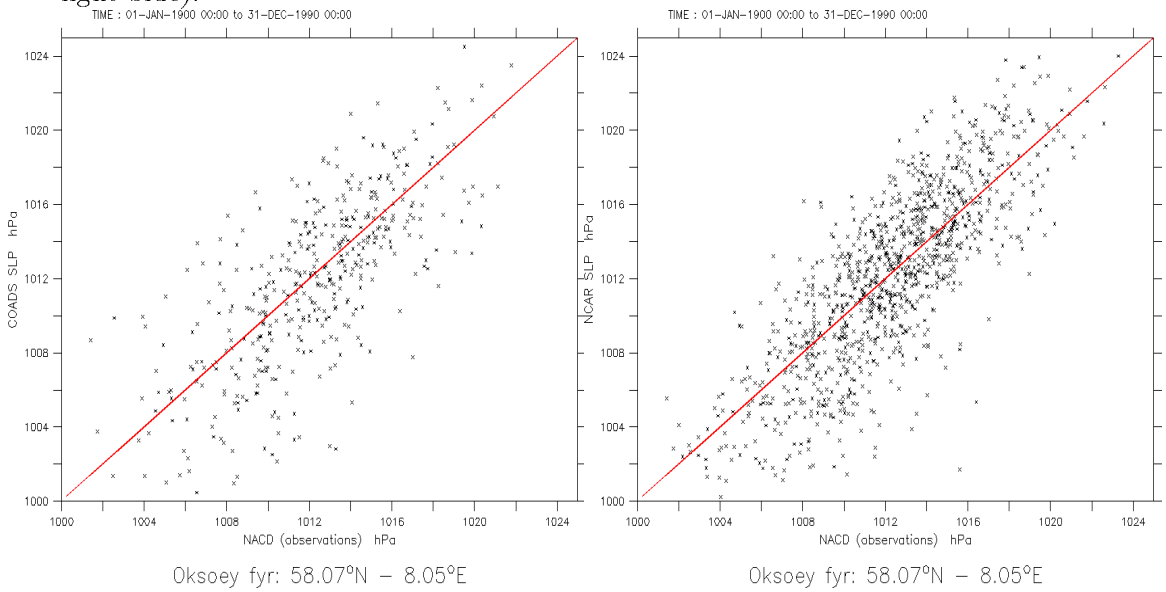

Figure 6: *Scatter plots of the SLP time series from Oksøy fyr show the bias and spread (error) of the data. The left panel shows the COADS SLP values plotted against the observations and the right panel shows the similar analysis for the NCAR ds010.0 data.*

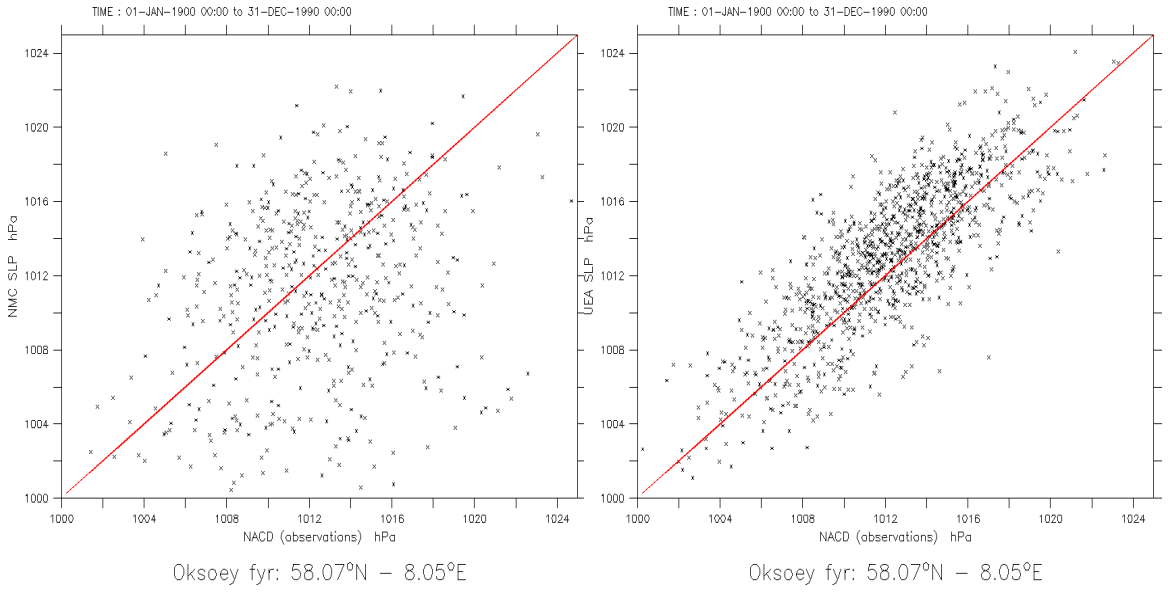

Figure 7: *Scatter plots of the SLP time series from Oksøy fyr show the bias and spread (error) of the data. The left panel shows the NMS ds195.5 SLP values plotted against the observations and the right panel shows the similar analysis for the UEA SLP data.*

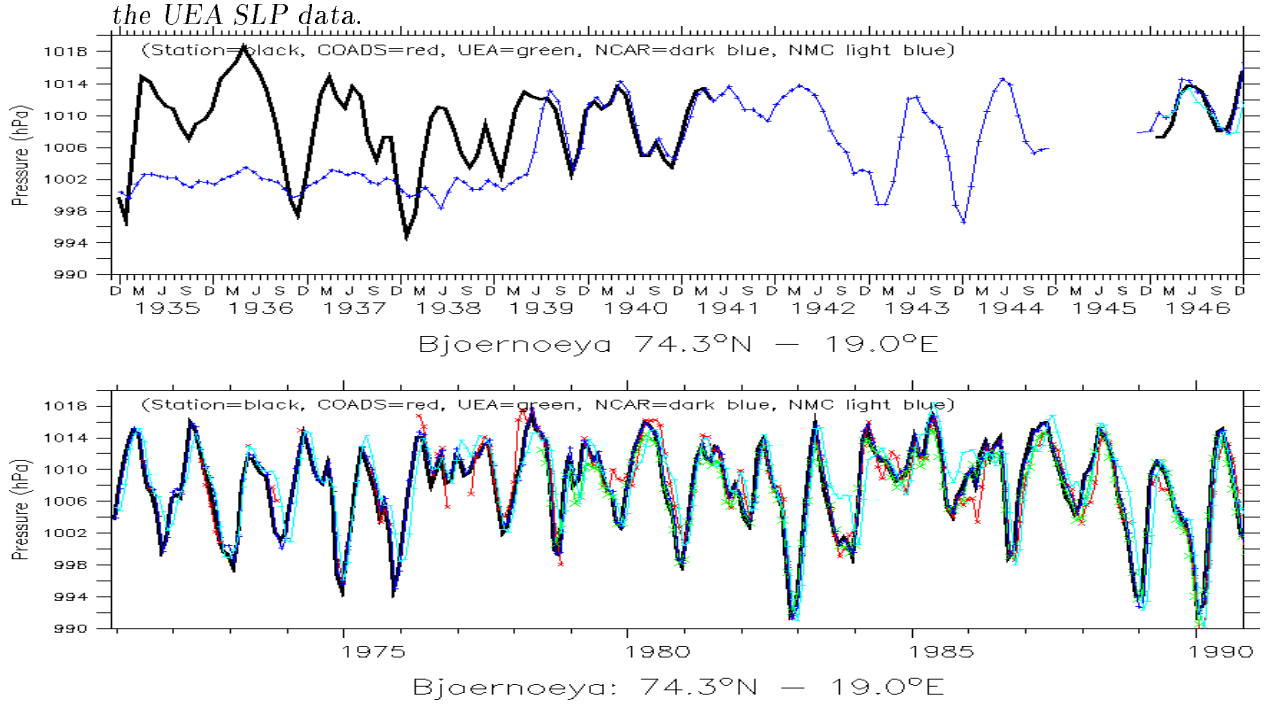

Figure 8: *Time series of monthly mean SLP values from Bjørnøya and from the predictor data sets (Station observation in black and COADS SLP in red; UEA MSLP in green; NCAR (ds010.0) in dark blue).*

pressure gradients will not be analysed here.

An assessment of the SST data sets was simply made by comparing the time series of area averages of the respective SSTs. Figure 7 shows the time series for the area 40°N-50°N/20°W-11°W, where the COADS SSTs were systematically lower than the GISST SSTs by almost 0.5°C before 1960. Apart from these systematic differences, the two data sets suggest similar fluctuations, with a general warming from early 20th century to 1955. A simple Monte Carlo resampling test revealed that this warming trend was outside the 2.5% - 97.5% confidence limits, and hence statistically significant.

3.6 Empirical Orthogonal Function analysis (PCA)

The Empirical Orthogonal Function (EOF) analysis was only applied to one calendar month at the time (i.e. only using January data). The principal components (lower panels figure 11) show little autocorrelation and suggest that the sub-sampled time series contained little red noise. The year-to-year fluctuations are strong compared with decadal variability.

The leading EOFs of the January mean SLPs for the UEA and NCAR data are shown in upper panel of figure 11. The data were not subject to geographical filtering prior to the PCA, and variability all over the northern hemisphere carried equal weight[11]. The PCA results indicate that variability over North Pacific is the most prominent feature in the SLP data. The first EOF patterns are different near northern Alaska and Siberia, where the NCAR SLPs indicate strong meridional gradients. The maximum over Iceland is displaced northward over towards Greenland in the NCAR data set, and the positive maximum over the Azores is weaker in the NCAR data set. Apart from these differences, the two EOFs are relatively similar, suggesting that the two data sets describe similar spatial structures of coherent variability. Both indicate strong variability over the North Pacific and an Atlantic north-south dipole pattern with maxima over Iceland and the Azores. This type of dipole pressure pattern over the North Atlantic implies enhanced westerly geostrophic winds for the British isles and Scandinavia. The eigenvalues of the EOFs are also similar for the two data sets, which means that the similar patterns in the two data sets describe similar variance. The regions near the Himalayas and latitudes north of 80°N in the NCAR data set were not included in the PCA due to problems with bad data (see section on the description of the NCAR data set). The UEA SLPs were masked north of 70°N in order to remove bad data.

The leading January EOF from the NMC ds195.5 data are compared with the corresponding UEA EOF in figure 12. Both EOFs show strong variability over Iceland and the Aleutian islands, as well as over North Pacific, however, there are some differences in the leading EOFs. The NMC EOF suggests similar strength

[11]The data was only weighted with the geometric weights: $cos\Phi$ prior to the analysis, where Φ is the latitude.

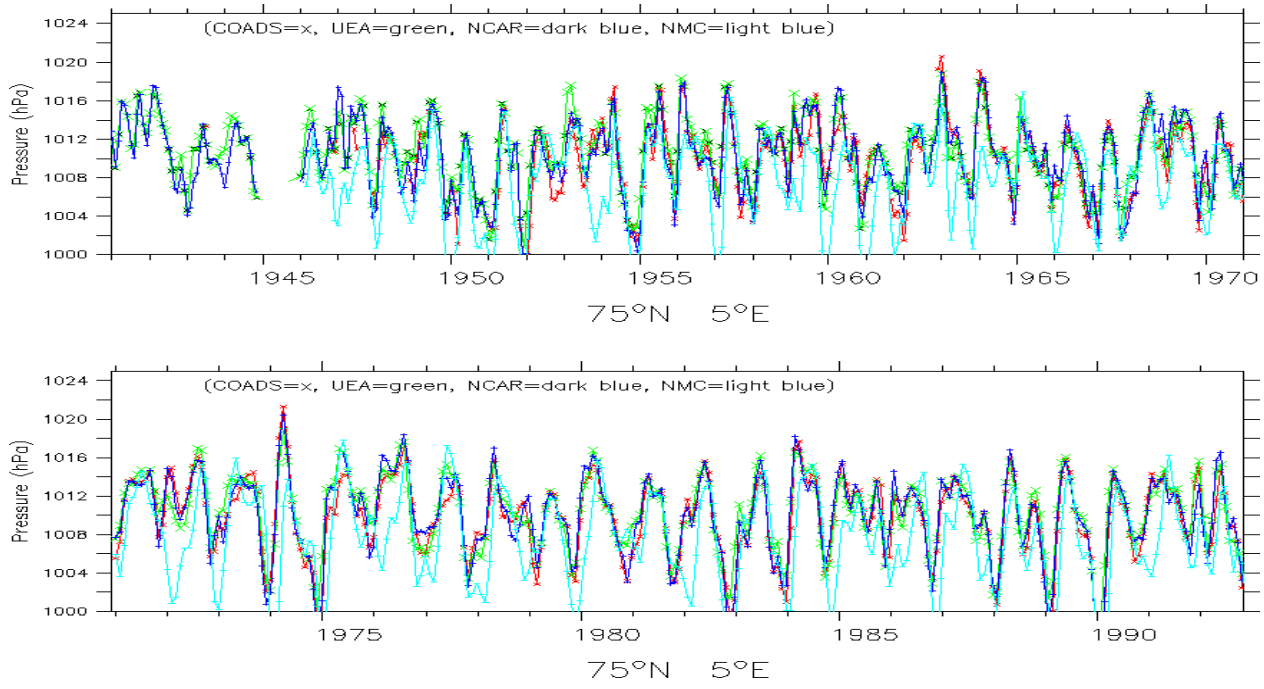

Figure 9: *Time series of monthly mean SLP values from 75°N , 5°E. (COADS SLP as 'x'; UEA SLP in green; NCAR (ds010.0) in dark blue; NMC (NCEP: ds195.5) in light blue).*

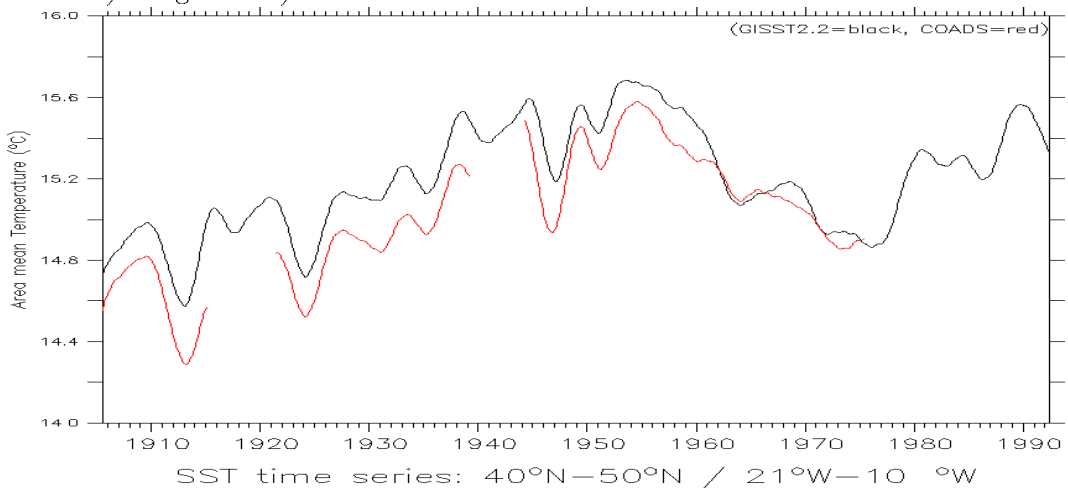

Figure 10: *Time series of spatial averages of SST: comparison between the GISST2.2 and COADS data sets. The data have been low pass filter by using a Hanning filter with a 61 month window width.*

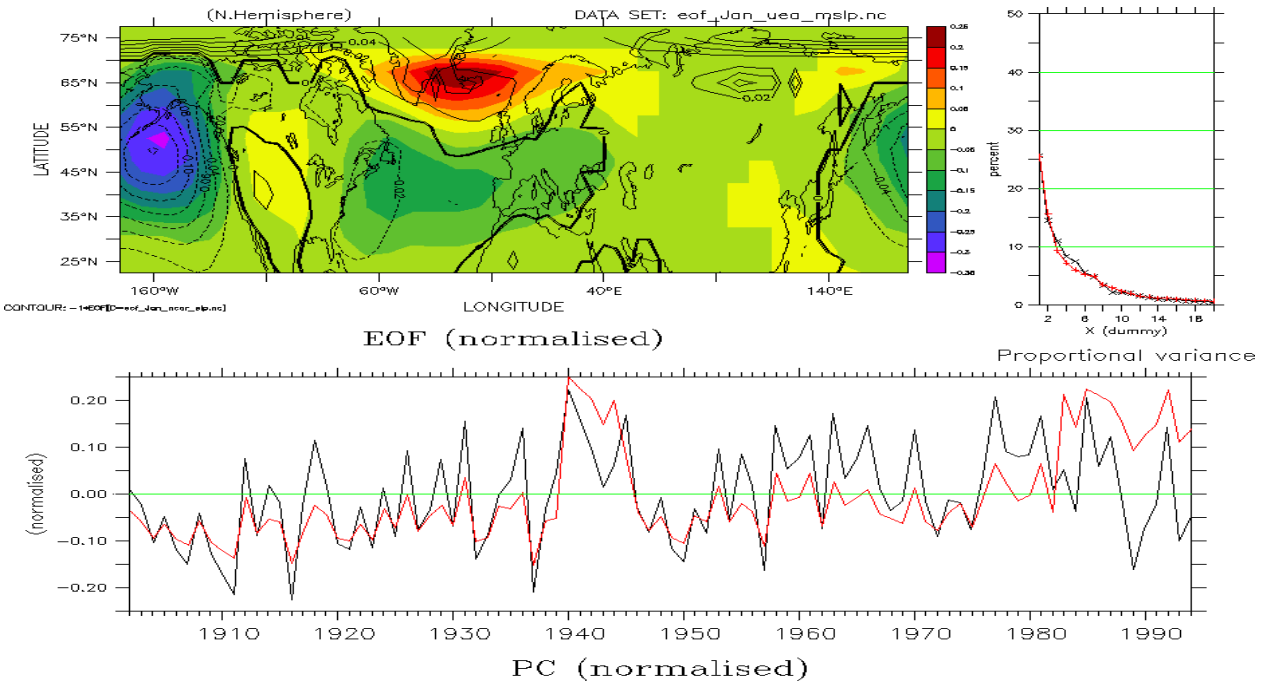

Figure 11: *Upper left panel shows the leading January EOF patterns for UEA (colours) and NCAR ds010.0 (contours) SLP. The right top panel shows the proportional variance described by the EOFs, or the "EOF eigenvalue spectrum". The bottom panel shows the time evolution of the leading EOFs, referred to as the principal components (PC).*

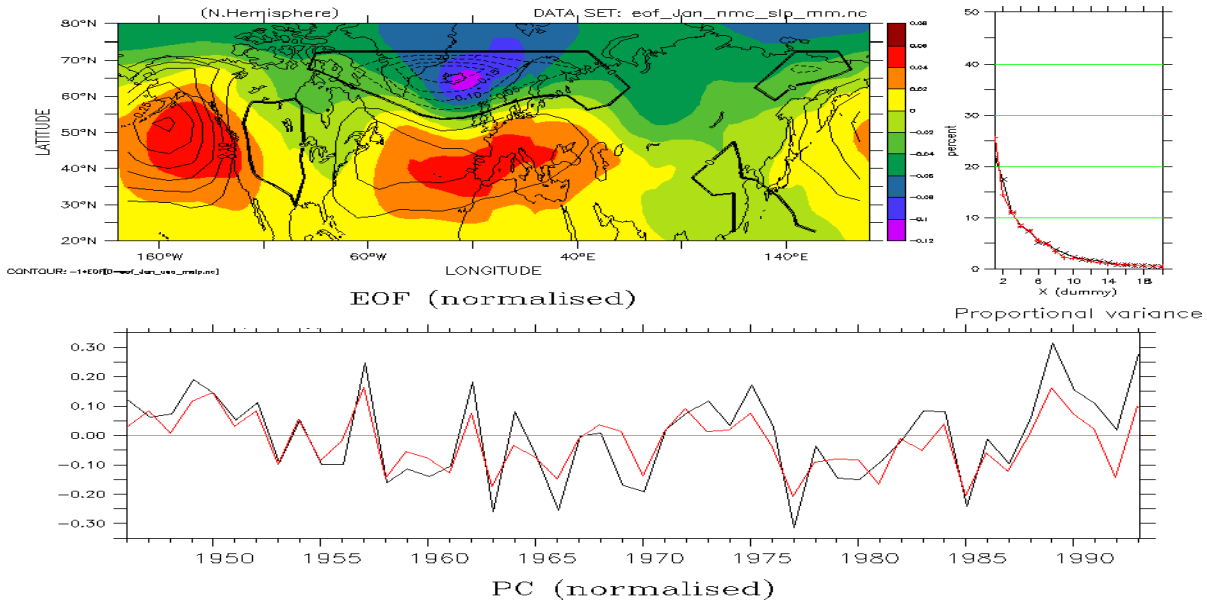

Figure 12: *Same as in figure 11, but for UEA and NMC SLP.*

of the SLP variability over the North Pacific and the Azores, but the UEA EOF has strongest loadings over the Aleutian islands. The UEA data set describes an Atlantic North-South dipole pattern with the southern maximum located over the Azores and the Atlantic ocean, while the NMC's southern maximum is over Iberia and France.

The regional differences in the EOFs over the North Atlantic were explored further by applying a geographical filter that emphasised the North Atlantic[12]. In this regional analysis, there was a better correspondence between the NCAR and UEA data, with both showing similar strength and locations for the North Atlantic dipole pattern (figure 13). The principal components of the leading UEA and NCAR EOFs indicate a high correlation. Figure 14 shows the North Atlantic PCA results for the UEA and NMC SLPs. The North Atlantic dipole pattern is again the most prominent feature, although the southern maximum of the NMC data is located further east than in the NCAR and UEA EOFs.

The PCA and CCA were also applied to the April, July, and October months, but during these months the regions south of 30°N contained some bad data in both the NCAR and the NMC data sets. Hence, the most southerly regions of the NCAR and NMC data sets were excluded from the analysis. Comparisons between the CCA results from the different data sets suggested similar spatial patterns, however, the PCA gave sometimes slightly different patterns (not shown). The locations of the North Atlantic dipole maxima were slightly different and the NCAR data described a much weaker pressure system over the Gulf of Alaska during April. Some of the July differences in the EOFs were associated regions of bad data over the middle East and Arctic in the NCAR data, however, main features such as the North Pacific pressure system and the North Atlantic dipole pattern were still present in the EOFs although these were not located in exactly the same regions. The NCAR leading EOF for October lacked a strong pressure system over the Gulf of Alaska, but the North Atlantic dipole pattern was present in all the PCA results. The EOF eigenvalues were similar for all the data sets and all the seasons.

The leading EOF for the GISST2.2 January SSTs is shown in figure 15. The main features of this EOF include a tripole pattern with positive maxima located in the Labrador sea and eastern tropical Atlantic, and with negative maximum just off the coast of New England. The principal component indicates that this mode is associated with both interannual as well as interdecadal time scales. This pattern is similar to the SST pattern that *Sutton & Allen* (1997) associated with a coupled ocean-atmosphere mode of variability. The SST variability described by this EOF is located near the storm track area, and, hence, may be important in terms of influencing the atmosphere. The eigenvalues corresponding to the leading EOF suggests that this EOF can account for 17% of the SST variance in the North Atlantic. The eigenvalues of the first and second EOFs are very close, and it is likely that these EOFs are degenerate (*North* et al., 1982).

[12] $xw(\theta) = |\sqrt{\cos(\theta)}|$, $yw(\Psi) = \cos(\Psi)\sqrt{\cos(\Psi - 60)}$.

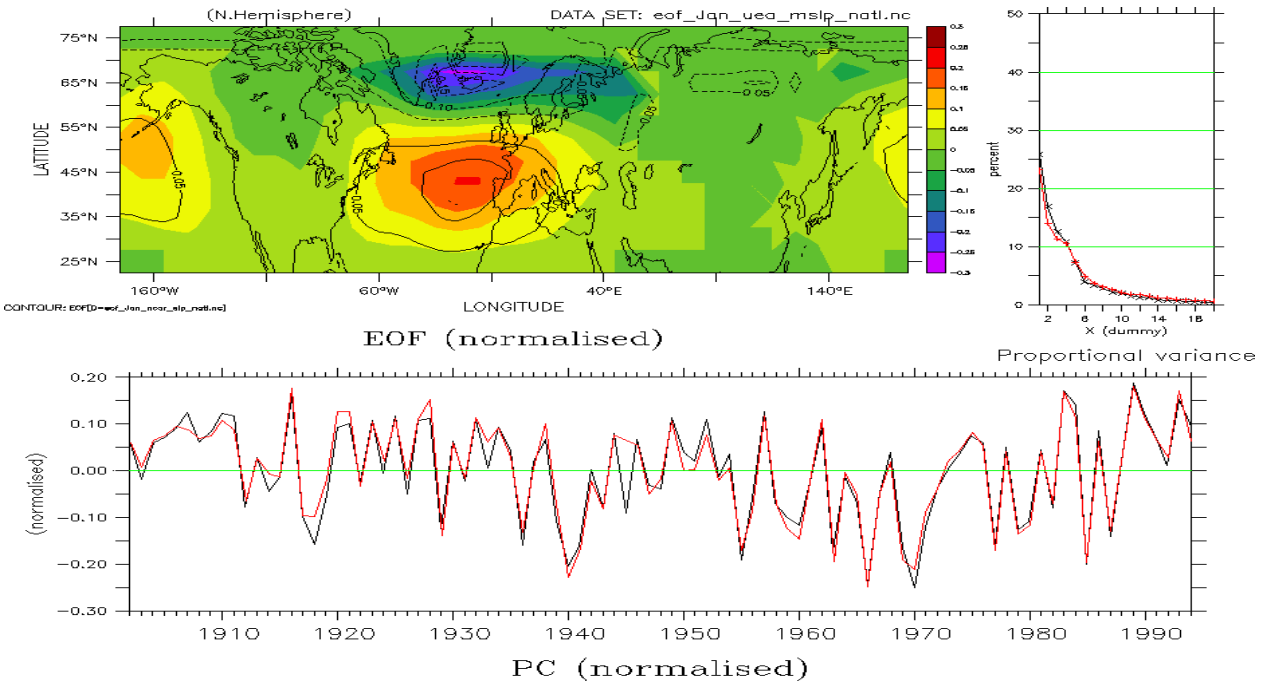

Figure 13: *Upper left panel shows the leading January EOF patterns for UEA SLP (colours) and NCAR ds010.0 (contours). A geographical (with emphasis on the North Atlantic)filter was applied to the data prior to the analysis. The right top panel shows the proportional variance described by the EOFs, or the "EOF eigenvalue spectrum". The bottom panel shows the time evolution of the leading EOFs, referred to as the principal components (PC).*

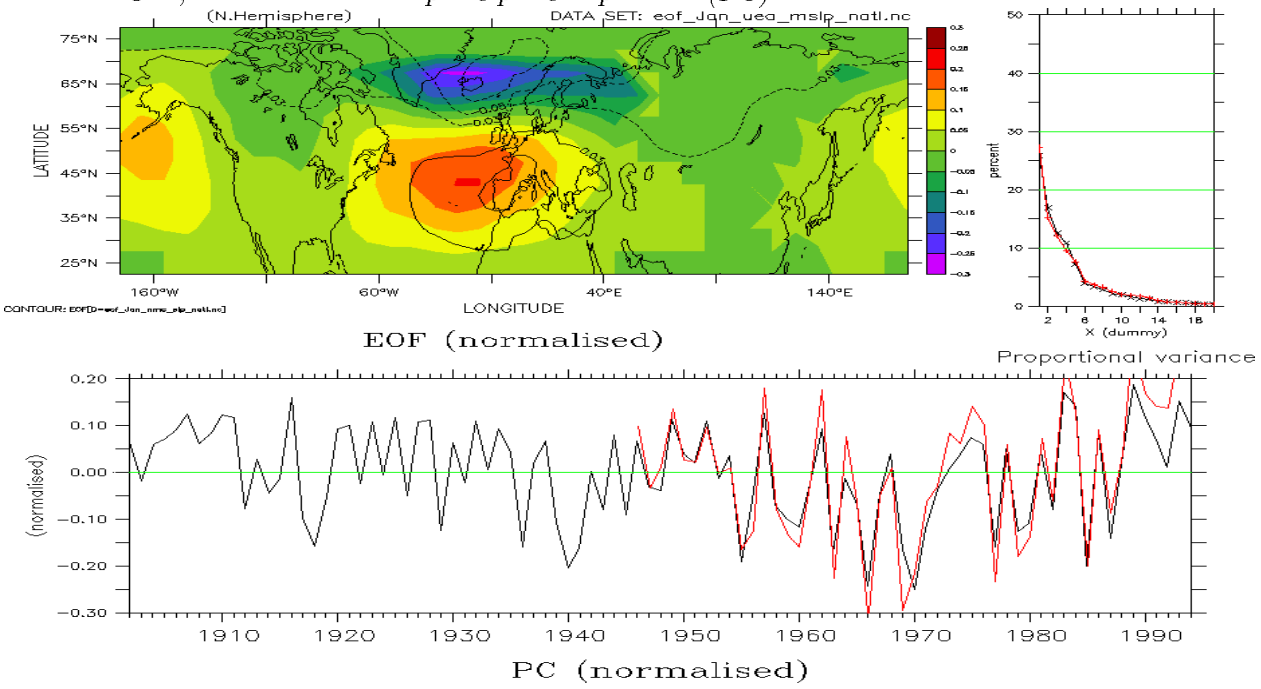

Figure 14: *Same as figure 13, but for UEA (colours) and NMC ds195.5 (contours) SLP.*

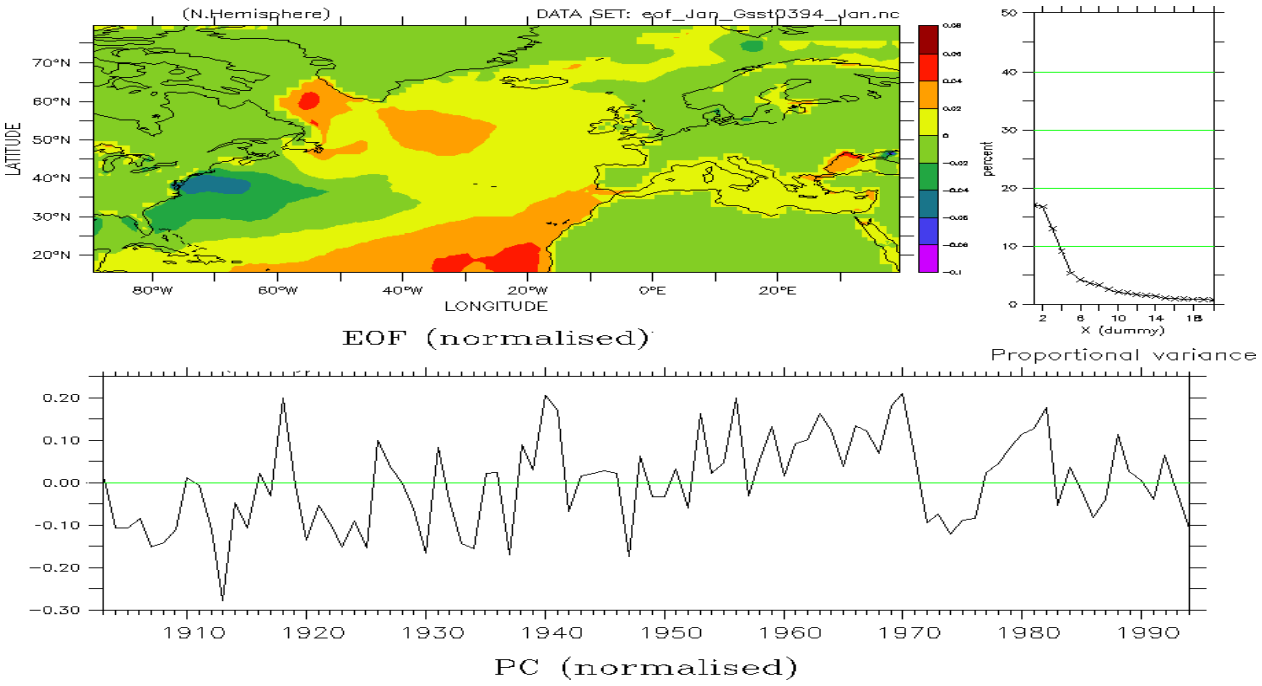

Figure 15: *The leading PCA results for the GISST2.2 January mean SSTs.*

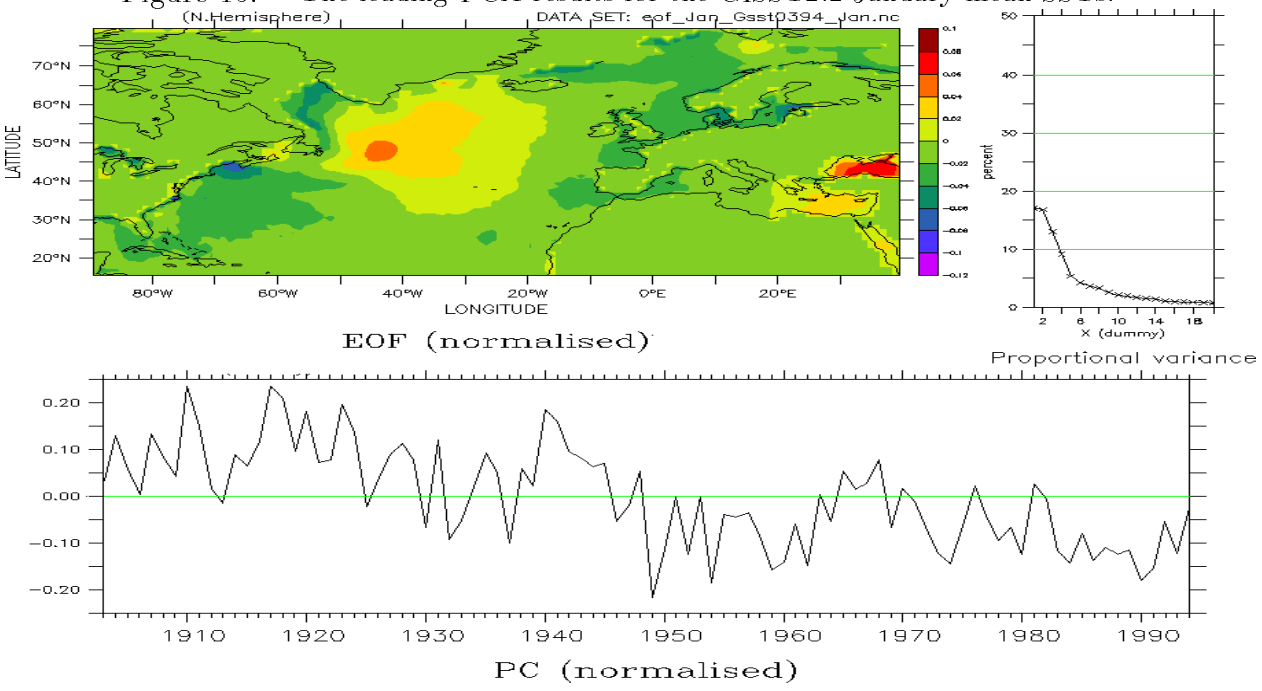

Figure 16: *The second PCA results for the GISST2.2 January mean SSTs.*

Figure 16 shows the second EOF of the GISST2.2 SSTs, indicating a dipole pattern with maxima near 50°N-40°W and just off the coast of New England. There are other regions such as the Black sea, Skagerak and Kattegat, as indicated by the leading EOF, that are coherent with the variability in the north-western Atlantic. It is not yet known whether there is a significant link between these different regions, but *Frankingnoul* (1985) have suggested that the mid latitude SST variability is a result of atmospheric forcing rather than advection by the ocean currents, which implies that large scale atmospheric circulation may influence SSTs over large areas. It is also possible that these features are artifacts of the EOF computation and hence coincidental.

The principal component suggests a slight long term trend, with cooling in the North sea between 1920 and 1970 and a slight warming south of Greenland. A simple Monte Carlo resampling test indicated that this trend was within the 2.5%-97.5% confidence limit and therefore *not* significant.

The third EOF (not shown) exhibits strong variability on decadal time scales near Cape Hatteras, where the Gulf stream separates from the coast. The time series describing the temporal evolution indicates variability on interdecadal time scales. The strong SST variability in this region may be associated with the northward and southwards excursions of the Gulf stream just off the east coast of America. The fact that the EOF pattern does not show a north-south dipole pattern in this region, however, suggests that the variability captured is not merely a result of the latitudinal displacement of the Gulf current.

The EOF analysis was applied to the Walsh January ice data from the GISST2.2 data set (*Rayner* et al., 1996), and the leading EOF is shown in figure 17. This EOF can account for a significant portion of the variance. The locations with most variability are along the ice edge of the Arctic and in the Baltic sea. The areas of strong variability along the ice edge have weights with the same sign, suggesting that the ice around the entire Arctic expands and retracts in phase. Some ice is also apparent in the northern end of the Caspian sea.

3.7 CCA

The CCA described here used data prefiltered through 20 EOFs (described in the section above) as input data. The approach is described in more detail by *Barnett & Preisendorfer* (1987) and in *Bretherton* et al. (1992).

Figure 18 shows the January CCA patterns (no geographical filtering was applied to the data prior to the analysis) for the UEA and NCAR ds010.0 SLP data sets. The two data sets show similar CCA patterns with minor differences. The maximum over the Aleutian islands is slightly more pronounced and the maximum over Iceland is located slightly further south and contains some bad data in the NCAR data (top left).

The CCA results for NMC and UEA reveals some important differences over the UK and the North Sea, where the maximum in this are extends further east

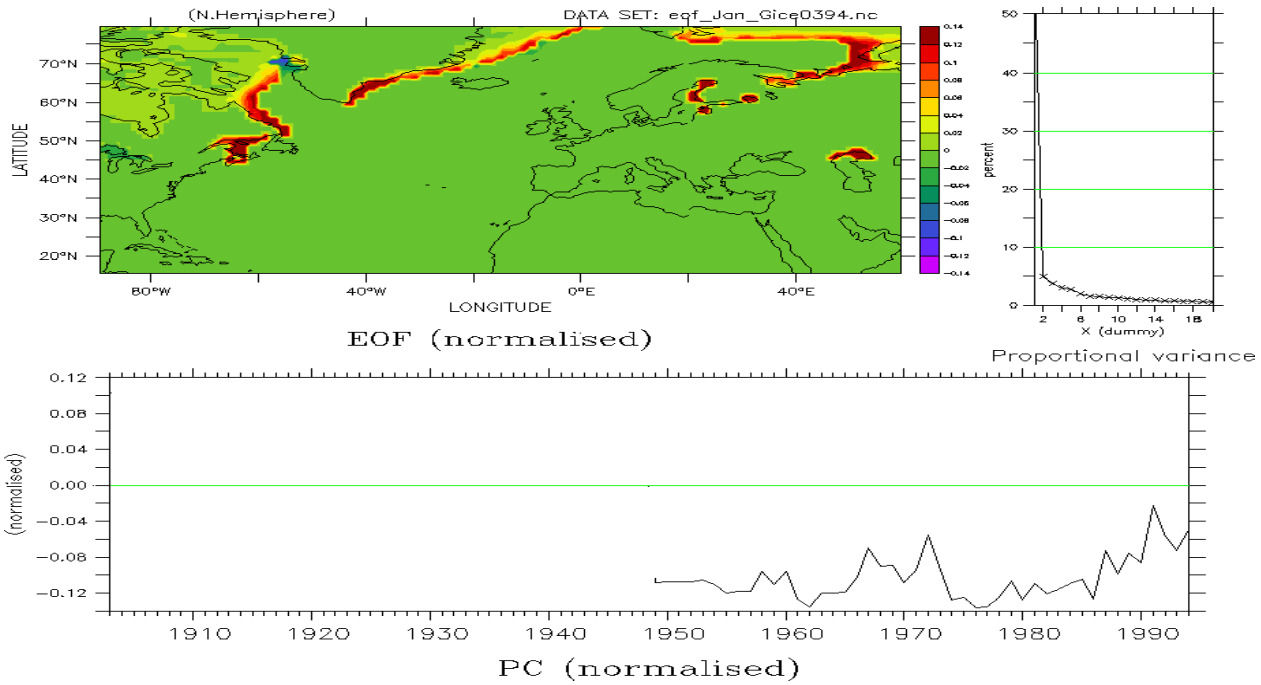

Figure 17: *The leading PCA results for the GISST2.2 January mean ice data.*

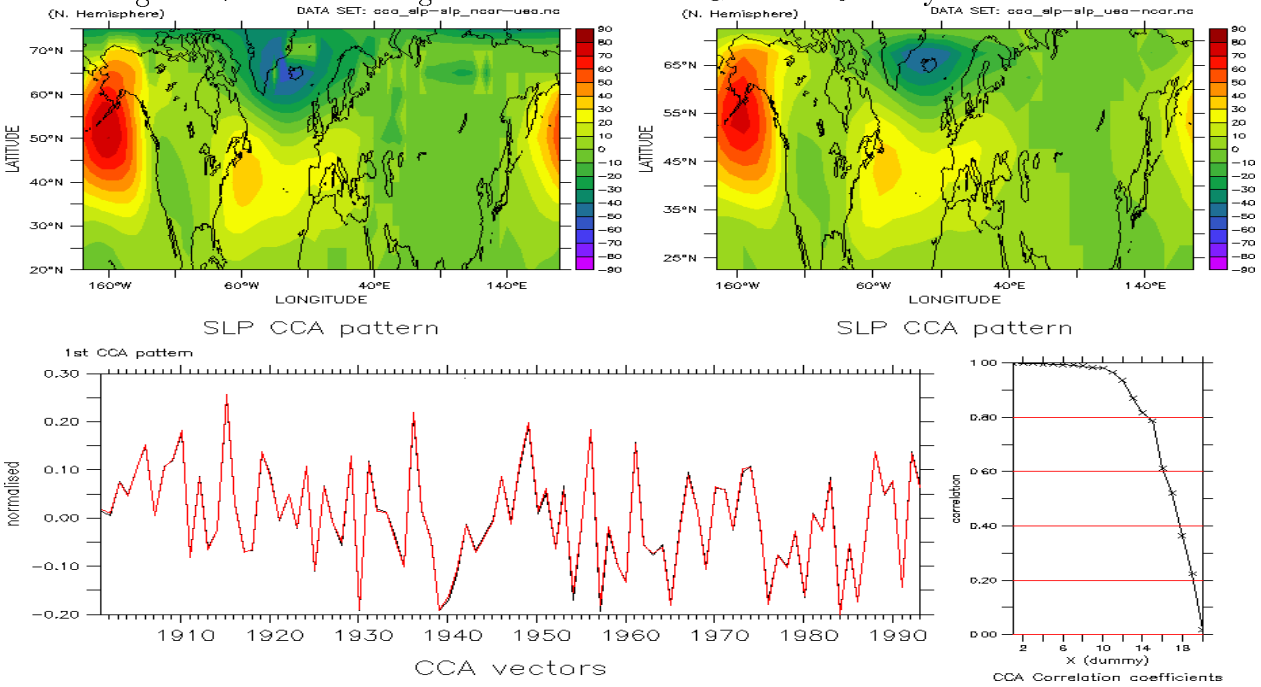

Figure 18: *The leading UEA-NCAR (ds010.0) CCA results for the January mean SLPs. The upper left panel shows the NCAR CCA pattern and the top right panel shows the corresponding UEA pattern. The lower panels show the canonical expansion coefficients (left) and the canonical correlations (right).*

in the NMC data set. The differences in the pressure patterns may imply different geostrophic wind, where for the NMC pattern Northern Norway implies mean winds from a more westerly direction than for the UEA pattern. There are also some subtle differences over eastern Russia and British Columbia, but the pattern are on the whole very similar.

The comparison between the NMC and NCAR CCA patterns indicates some differences over the Arctic seas. The NCAR data contains some bad data in the maximum over Iceland and Greenland (figure 20). The CCA patterns also reveals important differences over the Baltic countries, where the NMC pattern (right) suggests a northwesterly/southeasterly geostrophic flow that is absent in the NCAR data. The NCAR data furthermore suggests more westerly geostrophic flow over Southern Norway, whereas the NMC data describes the flow from a more southerly direction. The pressure system over the Aleutian islands is located further south-east in the NMC data.

4 Discussion

The intercomparison between the data sets revealed differences between the different gridded predictor data sets. If the data contained no errors, then these data sets are expected to be identical. The differences therefore give an indication of the accuracy of the data. Most of the differences between the data sets can be explained in terms of different analysis methods used for producing the data and different data sources. The time series comparison and the scatter plots indicate that the NMC (ds195.5) data account for more variance than the other data sets. The fact that the NMC grid has a higher spatial resolution may affect the variance and the EOF eigenvalue spectrum. In this case, this is unlikely as the NMC EOF eigenvalues are similar to those of the other data sets. One would expect the higher resolution data sets to contain more small scale noise that reduces the variance contribution from the leading EOFs.

The differences in the pressure patterns of the SLP data revealed by the CCA may have some implications for statistical downscaling. For instance, a CCA model trained on the NCAR January data will describe more southerly flow over Western Norway than a similar model trained on the NMC data because of the differences in the CCA SLP predictor patterns. A model that predicts the land surface temperatures may in this case 'think' that warmer temperatures over Western Norway are linked to westerlies if trained on the NMC data, or southerlies if trained on the NCAR data. This model may therefore predict slightly different temperatures for given a pressure pattern depending on which data was used for training. The spread in the predicted temperatures can be used to make a crude estimate of the uncertainty of the prediction.

It is difficult to say which data set is worse and which is better. Even the comparison with the station data may not tell which data set is closest to the truth, as these gridded data sets contain area mean values and not local values

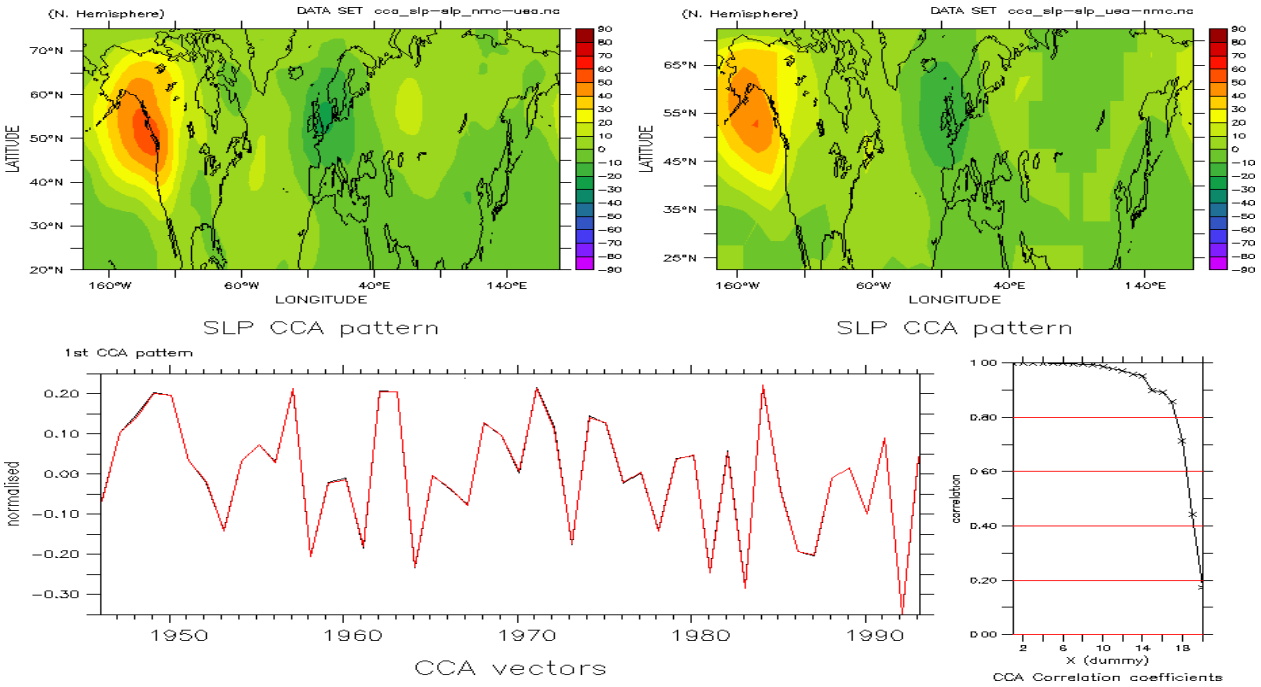

Figure 19: *The leading NMC-UEA CCA results for the January mean SLPs. The upper left panel shows the NMC CCA pattern and the top right panel shows the corresponding UEA pattern. The lower panels show the canonical expansion coefficients (left) and the canonical correlations (right).*

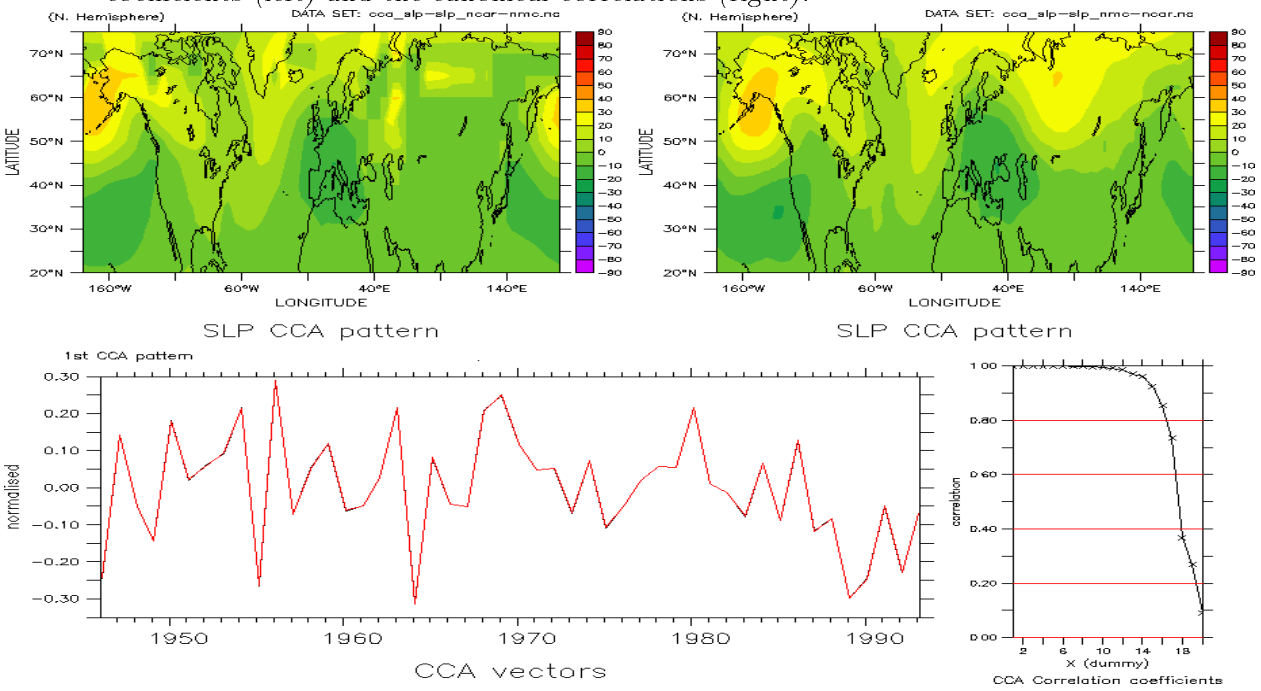

Figure 20: *The leading NMC-NCAR CCA results for the January mean SLPs. The upper left panel shows the NMC CCA pattern and the top right panel shows the corresponding NCAR pattern. The lower panels show the canonical expansion coefficients (left) and the canonical correlations (right).*

that may be affected be regional topography. The differences described above illustrates how difficult it is to obtain an objective and true description of the state of the climate. Using various data sets, however, gives a crude measure of uncertainty that can be attributed to the errors in the observations. The data coverage in the high latitudes is sparse before 1940, and therefore there are few long sequences of good quality data north of 70°N in any of the gridded data sets. Some of the bad data have affected the PCA and CCA results when the Arctic regions have been included, and the patches of bad data north of Iceland and along the east coast of Greenland in the NCAR data set, for instance, may produce errors in the predictions.

Finally, comparison between analyses of different data sets can provide an extra quality control since the results are expected to be similar for the different data sets. The intercomparison between the different data sets has been used to detect bugs in the conversion to the *netCDF* format. In our case, there were two instances of corrupt data which at first sight looked all right, as both the NCAR and the NMC data sets were stored in unfriendly and difficult formats and the FORTRAN codes provided to were hard to follow. The intercomparison revealed a small systematic phase shift in the NCAR data which caused by a bug, which otherwise would easily have been overlooked.

References

Barnett, T.P., & Preisendorfer, R.W. Origins and levels of monthlu and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review*, **115**, 1825–1850. 1987.

Basnett, T.A., & Parker, D.E. *Development of the Global Mean Sea Level Pressure Data Set GMSLP2.* Climatic Research Technical Note 79. Hadley Centre, Meteorological Office, Bracknell. 1997.

Bretherton, C.S, Smith, C., & Wallace, J.M. An Intercomparison of Methods for finding Coupled Patterns in Climate Data. *Journal of Climate*, **5**, 541–560. 1992.

Frankingnoul. Sea Surface Temperature Anomalies, Planetary Waves, and Air-Sea Feedback in the Mid Latitudes. *Review of Geophysics*, **23**(4), 357–390. 1985.

Jones, P.D. The early twentieth century Arctic High - fact or fiction? *Climate Dynamics*, **1**, 63–75. 1992.

NMC. *National Meteorological Center Grid Point Data set, CDROM: Version III, General Information and User's Guide.* Department of Atmospheric Sciences, Univeristy of Washington and Data Support Section, National Center for Atmospheric Research. 1996 (June).

North, G.R., Bell, T.L., & Cahalan, R.F. Sampling Errors in the Estimation of Emperical Orthogonal Functions. *Monthly Weather Review*, **110**, 699–706. 1982.

Parker, D.E., Jackson, M., & Horton, E.B. *The 1961- 1990 GISST2.2 Sea Surface Temperature and Sea-Ice Climatology.* Climate Research Technical Note 63. Hadley Centre, Meteorological Office, Bracknell. 1995.

Rayner, N.A., Horton, E.B., Parker, D.E., Folland, C.K., & Hackett, R.B. *Version 2.2 of the Global sea-Ice and Sea Surface Temperature data set, 1903-1994.* Climate Research Technical Note 74. Hadley Centre, Meteorological Office, Bracknell. 1996.

Slutz, R.J., Lubker, S.J., Hiscox, J.D., Woodruff, S.D., Jenne, R.L., Steurer, P.M., & Elms, J.D. *Comprehensive Ocean-Atmosphere Data Set; Release 1.* Tech. rept. Climate Research Program, Boulder, Colorado. 1985.

Sutton, R.T., & Allen, M.R. Decadal predictability of North Atlantic sea surface temperature and climate. *Nature*, **388**, 563–567. 1997.

Trenberth, K.E., & Paolino, D.A. The Northern Hemisphere sea-level pressure data set: trends, errors and discontinuities. *Monthly Weather Review*, **104**, 1354–1361. 1980.

Wilks, D.S. *Statistical Methods in the Atmospheric Sciences.* Orlando, Florida, USA: Academic Press. 1995.

5 Appendix: list over data sources

(The list is a copy of RegClim PT3's local intranet page at the DNMI, and text is therefore in Norwegian.)