CHAPTER 3: METHODS AND ANALYSES APPLIED.

3.1. INTRODUCTION.

This chapter deals with the extraction of the meteorological variables to be used in the analyses and the description of different mathematical and statistical tools to explore the patterns and variability of the climatic change in Mexico.

The construction of a large network of long-term high-quality databases of daily precipitation and temperature is addressed in the first part of the chapter. The extraction procedure for these meteorological time-series and the process of data quality control are both explained here. Because one of the purposes of the thesis is to link climate change patterns in Mexico during instrumental periods with the El Niño-Southern Oscillation (ENSO) phenomenon, the extraction of the three different indices (SOI, Niño 3.4 and MEI) used in these thesis are also considered in the first half of this chapter.

Three main mathematical methods are discussed in the second half of this chapter. The first is the application of Principal Component Analysis as a tool to find groups of stations that vary coherently, together with their use in calculating weighted regional averages. The second topic deals with the changes of the climatic variables at the fringes of their probability distributions, usually called weather extremes. The last method describes the two different approaches to estimate correlations between meteorological variables. Non-parametric correlations are obtained using Kendall's tau, as an alternative (to measure the association between the time-series) to the extensively used linear correlation is shown first. Lag-cross correlations are finally presented as a tool to find the lag that maximises the coherence (linear correlation) between a pair of variables.

3.2. DATA EXTRACTION.

In México the longest meteorological time series are those of (land surface) precipitation and air temperature, especially the former. This is true for either daily or monthly data. As several studies have been made using monthly values, daily figures were the first objective of the extraction, in order to explore the possibility of having a database of relatively long climatic records with high temporal resolution.

3.2.1. PRECIPITATION AND TEMPERATURE DAILY DATA.

Among the digitized data considered (because of their digital accessibility and length) are:

DAT322©

This software was prepared by the Mexican Institute of Water Technology (Instituto Mexicano de Tecnología del Agua, IMTA) to manage the 322 meteorological stations with the longest time series. The selection of the stations included was made by the Mexican Meteorological Office (Servicio Meteorológico Nacional, SMN). The documentation of this software states that climatological analyses were performed according to the Manual of the CLImat COMputing (CLICOM) project of the World Meteorological Organization (WMO) to identify outliers in the information. Strangely, it does not contain several of the largest cities (presumably with the longest data files) in Mexico, failing to present a complete national picture of the potential instrumental records. Another problem found is that missing values are defined with a zero value instead of the other options conventionally accepted.

ERIC©

This software was also prepared by IMTA with the latest version released in 2000, and contains daily precipitation and temperature data among other variables. Most of the stations have information from 1960 to 1995. No data quality analyses were performed in this database, and being typed manually this is not a minor issue. For instance, for certain months at several stations, temperature data was typed instead of precipitation in

the rainfall time series. That is why, careful attention and reserved use was given to this source.

CLICOM

Another source of data is the already mentioned CLImat COMputing Project (CLICOM) of the World Meteorological Organization (WMO). It incorporates digital daily data for almost all the stations considered by DAT322 and ERIC, but some of them have been updated until 2002 inclusive. Like DAT 322 this database does not include sufficient information for the largest - and most of the times oldest - cities in Mexico.

GASIR

GASIR (Gerencia de Aguas Superficiales e Ingeniería de Ríos) was developed by the office of Dams operation and river engineering of the National Water Commission (Comisión Nacional del Agua, CNA). They received daily precipitation data from many stations located across the country. Unfortunately, they only have digital information available from 1989 to 2001, so this database was used mainly to complete many recent gaps. Because this source is used mainly for reservoir purposes, its format is slightly different (the date is one day ahead) from the other databases, a program in fortran was needed to adapt the precipitation values to the WMO general rules. An important aspect of GASIR data, on the other hand, is that it has good quality as a whole and is almost free of errors.

3.2.2. STATIONS SELECTED.

Having all those daily digital databases available, it was necessary to choose the most appropriate stations for the subsequent analyses. Because, according to its documentation, DAT322 claimed to have the longest records selected by SMN, it was selected as the first reference or the start of the extraction for every station to be processed. The first condition defined – for climatological reasons - was that every station to be considered should have at least thirty years of information. Less than ten per cent of missing values was considered as a second limitation to extract a time series. So,

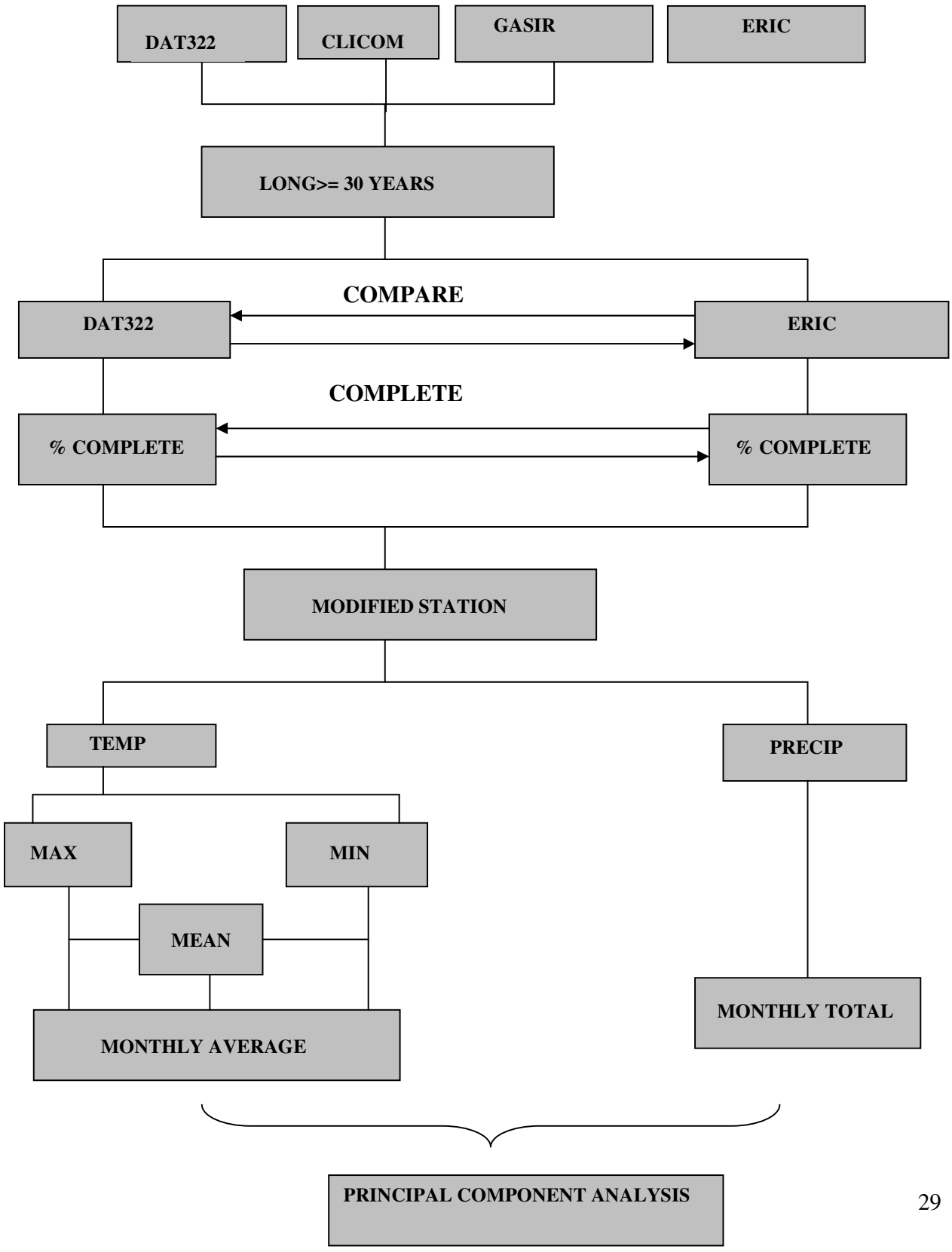
every other source already mentioned with time series fulfilling both conditions was included initially.

DATA EXTRACTION PROCEDURE.

Having enough information is not a sufficient condition for climatic studies. It is essential to assure the quality of the data extracted. That is why a procedure was designed to compare and complete the data for every station to be processed. Basically, it could be described as follows: every station in DAT322 having thirty or more years of information was compared with the other different sources and the missing values filled when possible. Then, with the daily data ready, a process to compute monthly values (mean temperatures or accumulated precipitation) and basic statistics was performed. The maximum number of missing values allowed for a month was set to four, otherwise the month was considered as missing. With these statistics it was also possible to identify (for instance, in comparison with known climatological normals) suspicious values (see fig. 3.1). An example in which data from ERIC substitute missing values (tagged as -1 in fig. 3.2) in the CLICLOM database. Another case in which the values in the CLICOM database has been multiply by a factor of ten are replaced with the ERIC data (see fig. 3.3).

After all this information was processed, only a limited number (93 stations, see table 3.1 and fig. 3.4) of daily data stations were considered as being long enough, resulting - already pointed out by several authors - in the sparsity of the meteorological observations network (O' Hara and Metcalfe, 1995; Englehart and Douglas, 2003). So, another source was used. Such a database is a monthly precipitation collection from 1931 to 1989; it was prepared by Carlos Espinosa Cruishank (specialist in Hydraulics) in the SMN. Hence, a triple checking process was made with every time series: among Espinosa's monthly data, climatological monthly figures by García (1988), and the data processed (DAT322, CLICOM, ERIC, and GASIR) from the stations reporting daily. Finally, a plot of every annual time series was made in order to find any inconsistency among of them.

Fig. 3.1. DAILY DATA EXTRACTION PROCEDURE



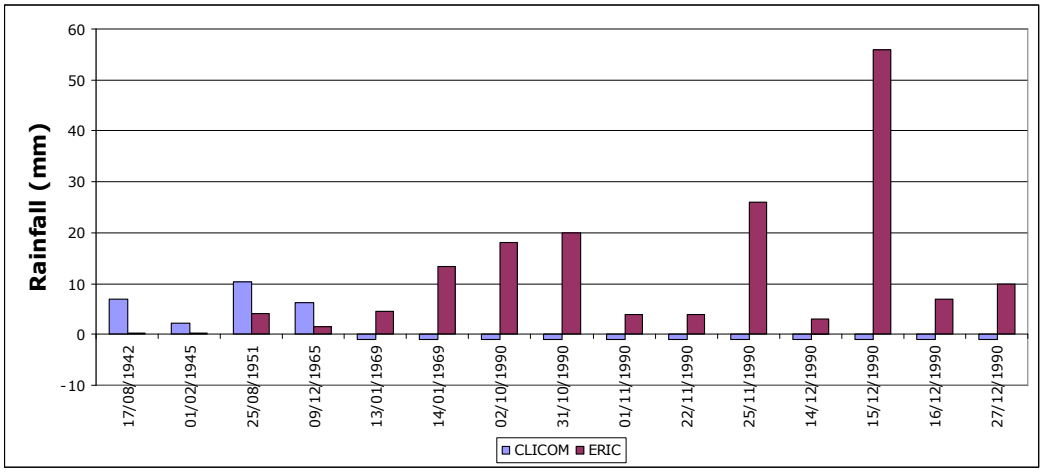


Fig. 3.2. Example of daily precipitation data being corrected. Two different databases are compared. In this case, missing values (-1) in one time-series (CLICOM) are corrected with the second dataset (ERIC).

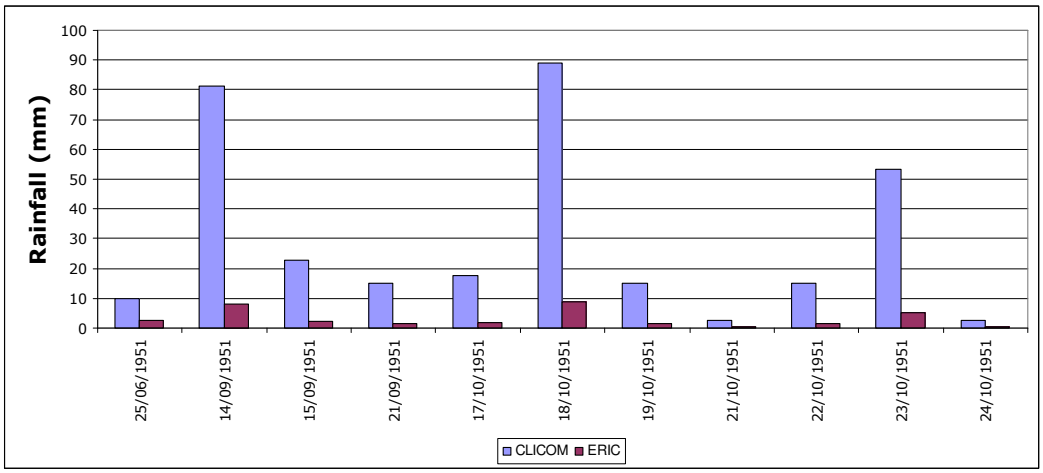


Fig. 3.3. Example of daily precipitation data being corrected. Two different databases are compared. In this case, a systematic error (values are multiplied by a factor of 10) in the first time-series (CLICOM) are substituted by the corrected values in the second dataset (ERIC).

	STATION NAME	STATE	SMN ID	LONGITUDE	LATITUDE	ALTITUDE*
1	PABELLON DE ARTEAGA	AGS	01014	-102.33	22.18	1920
2	PRESA CALLES	AGS	01018	-102.43	22.13	2025
3	PRESA RODRIGUEZ	BCN	02038	-116.90	32.45	100
4	EL PASO DE IRITU	BCS	03012	-111.12	24.77	140
5	LA PURíSIMA	BCS	03029	-112.08	26.18	95
6	LORETO	BCS	03035	-111.33	26.00	15
7	SAN JOSE DEL CABO	BCS	03056	-109.67	23.05	7
8	SANTA ROSALíA	BCS	03061	-112.28	27.30	17
9	SANTIAGO	BCS	03062	-109.73	23.47	125
10	TODOS SANTOS (DGE)	BCS	03066	-110.22	23.43	18
11	CHAMPOTON	CAMP	04008	-90.72	19.35	2
12	HECELCHAKAN	CAMP	04011	-90.13	20.18	13
13	SABANCUY	CAMP	04029	-91.11	18.97	2
14	RAMOS ARIZPE	COAH	05032	-100.98	25.53	1399
15	SALTILLO	COAH	05048	-101.00	25.42	645
16	COLIMA	COL	06040	-103.73	19.23	495
17	OCOZOCUAUTLA	CHIAP	07123	-93.38	16.70	864
17	CIUDAD GUERRERO	CHIH	08028	-108.52	28.52	2000
18	CD CUAUHTEMOC	CHIH	08026	-106.85	28.42	2050
19	CIUDAD DELICIAS	CHIH	08044	-105.43	28.20	1170
20	HIDALGO DEL PARRAL	CHIH	08078	-105.67	26.93	1950
21	LA JUNTA	CHIH	08090	-107.97	28.75	1900
22	BATOPILAS	CHIH	08161	-107.75	27.02	556
23	EL PALMITO	DUR	10021	-104.78	25.52	1540
24	FCO. I MADERO	DUR	10027	-104.30	24.47	1960
25	GUANACEVI	DUR	10029	-105.97	25.93	2200
26	RODEO	DUR	10060	-104.53	25.18	1340
27	SAN MARCOS	DUR	10070	-103.50	24.27	
28	SANTIAGO PAPASQUIARO	DUR	10100	-105.42	25.05	1740
29	IRAPUATO	GTO	11028	-101.35	20.68	1725
30	SAN DIEGO DE LA UNION	GTO	11064	-100.87	21.47	2080
31	SAN JOSE ITURBIDE	GTO	11066	-100.40	21.00	2100
32	AYUTLA (CFE)	GRO	12012	-99.10	16.95	
33	CHILAPA	GRO	12110	-99.18	17.60	1450
34	HUICHAPAN	HGO	13012	-99.65	20.38	1102
35	MIXQUIHUALA	HGO	13018	-99.20	20.23	2050
36	CHAPALA	JAL	14040	-103.20	20.30	1523
37	MASCOTA	JAL	14096	-104.82	20.52	1240
38	SAN FRANCISCO	MEX	15089	-99.97	19.30	2630
39	LA PIEDAD CABADAS (DGE)	MICH	16065	-102.03	20.37	1700
40	TACAMBARO	MICH	16123	-101.47	19.23	1820
41	YURECUARO	MICH	16141	-102.28	20.35	1537
42	ZAMORA	MICH	16144	-102.28	20.00	1540
43	CUERNAVACA	MOR	17004	-99.25	18.92	1529
44	ACAPONETA	NAY	18001	-105.37	22.50	22
45	CADEREYTA	NL	19008	-100.00	25.60	350
46	EL CUCHILLO	NL	19016	-99.25	25.73	145

	STATION NAME	STATE	SMN ID	LONGITUDE	LATITUDE	ALTITUDE*
47	LOS RAMONES	NL	19042	-99.63	25.70	210
48	MONTEMORELOS	NL	19048	-99.83	25.20	425
49	MONTERREY	NL	19052	-100.30	25.68	540
50	JUCHITAN	OAX	20048	-95.03	16.43	46
51	MATIAS ROMERO	OAX	20068	-95.03	16.88	201
52	SANTO DOMINGO TEHUANTEPEC	OAX	20149	-95.23	16.33	95
53	PIAXTLA	PUE	21063	-98.25	18.20	1155
54	TEZIUTLAN	PUE	21091	-97.35	19.82	2050
55	HUAUCHINANGO	PUE	21118	-98.05	20.18	1575
56	JALPAN	QRO	22008	-99.47	21.22	860
57	PRESA CENTENARIO	QRO	22025	-99.90	20.52	1880
58	ALVARO OBREGON	QROO	23001	-88.62	18.30	
59	CHETUMAL	QROO	23032	-88.30	18.50	6
60	CHARCAS	SLP	24010	-101.12	23.13	2020
61	MATEHUALA	SLP	24040	-100.63	23.65	1575
62	MEXQUITIC	SLP	24042	-101.12	22.27	2030
63	SAN LUIS POTOSI (DGE)	SLP	24069	-100.97	22.15	1870
64	CIUDAD DEL MAIZ	SLP	24116	-99.60	22.40	1245
65	BADIRAGUATO	SIN	25110	-107.55	25.37	230
66	QUIRIEGO	SON	26075	-109.25	27.52	521
67	TRES HERMANOS	SON	26102	-109.20	27.20	100
68	YECORA	SON	26109	-108.95	28.37	
69	SAN FERNANDO	TAM	28086	-98.15	24.85	43
70	TAMPICO (DGE)	TAM	28111	-97.87	22.22	12
71	VILLAGRAN	TAM	28118	-99.48	24.48	380
72	APIZACO	TLX	29002	-98.13	19.42	2404
73	TLAXCALA	TLX	29030	-98.23	19.32	2552
74	TLAXCO	TLX	29032	-98.13	19.63	2444
75	CATEMACO	VER	30022	-95.10	18.42	338
76	CHICONTEPEC	VER	30041	-98.17	20.98	595
77	IXHUATLAN	VER	30072	-98.00	20.70	306
78	JALTIPAN	VER	30077	-94.43	17.97	46
79	PAPANTLA	VER	30125	-97.32	20.45	298
80	RINCONADA	VER	30141	-96.55	19.35	313
81	SOLEDAD DOBLADO	VER	30163	-96.42	19.05	183
82	VERACRUZ	VER	30192	-96.13	19.20	16
83	JALAPA	VER	30228	-96.92	19.53	1999
84	TUXPAN	VER	30229	-97.40	20.95	4
85	PANUCO	VER	30285	-98.17	22.05	60
86	PROGRESO	YUC	31023	-89.65	21.28	8
87	SOTUTA	YUC	31030	-89.02	20.60	11
88	MERIDA (DGE)	YUC	31044	-89.62	20.98	9
89	EL SAUZ	ZAC	32018	-103.23	23.18	2100
90	SOMBRERETE	ZAC	32054	-103.63	23.63	2300
91	JUCHIPILA	ZAC	32067	-103.13	21.42	1240
92	TEUL DE GLEZ. ORTEGA	ZAC	32070	-103.47	21.47	1900
93	ZACATECAS	ZAC	32086	-102.57	22.77	2450

Table 3.1. Spatially incomplete network of daily data stations for precipitation. The period of records for all the stations is from 1931 to 2001. * meters above sea level.

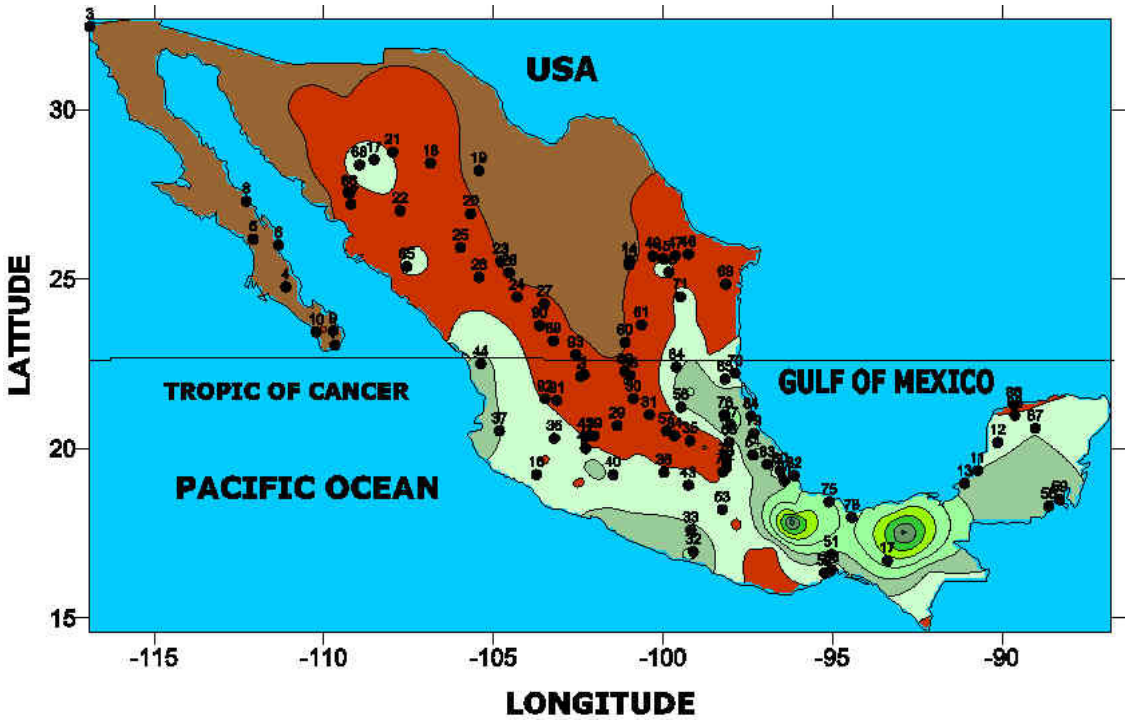


Fig. 3.4. Resulting network of 93 stations after the first stage of extraction of daily rainfall data.

The final network consists of a set of 175 stations having monthly precipitation, 168 are Mexican and 7 are southern USA stations, with good spatial coverage (Table 3.2, and Fig. 3.5). The length of every time series is of 71 years, starting in 1931 and ending in 2001. The maximum percentage of missing values was restricted to 10%.

It is possible that some information has been left out of the extraction efforts as there are some records still on paper in SMN, but this is unlikely to happen in terms of digital databases. All the currently known Mexican climatological digitised sources were considered. That is why an acceptable spatial coverage is expected, markedly better than all the few former studies aiming at a national appraisal of the Mexican climate using the longest time series available. For extraction purposes of precipitation and temperature data, the definitions of wet and dry seasons established in section 2.2.1 are applied in this chapter.

	STATION NAME	STATE	SMN ID	LONGITUDE	LATITUDE*	ALTITUDE+
1	AGUASCALIENTES	AGS	01001	-102.30	21.88	1870
2	PABELLON DE ARTEAGA	AGS	01014	-102.33	22.18	1920
3	PRESA CALLES	AGS	01018	-102.43	22.13	2025
4	PRESA RODRIGUEZ	BCN	02038	-116.90	32.45	100
5	ENSENADA	BCN	02072	-116.60	31.88	24
6	BUENAVISTA	BCS	03004	-111.80	25.10	30
7	EL PASO DE IRITU	BCS	03012	-111.12	24.77	140
8	LA PURíSIMA	BCS	03029	-112.08	26.18	95
9	LORETO	BCS	03035	-111.33	26.00	15
10	MULEGE	BCS	03038	-111.98	26.88	35
11	SAN BARTOLO	BCS	03050	-109.85	23.73	395
12	SAN JOSE DEL CABO	BCS	03056	-109.67	23.05	7
13	SANTA GERTRUDIS	BCS	03060	-110.10	23.48	350
14	SANTA ROSALíA	BCS	03061	-112.28	27.30	17
15	SANTIAGO	BCS	03062	-109.73	23.47	125
16	TODOS SANTOS (DGE)	BCS	03066	-110.22	23.43	18
17	LA PAZ	BCS	03074	-110.37	24.15	10
18	SABANCUY	CAMP	04029	-91.11	18.97	2
19	CAMPECHE	CAMP	04038	-90.53	19.85	8
20	CHAMPOTON	CAMP	04041	-90.72	19.37	2
21	PRESA VENUSTIANO CARRANZA	COAH	05030	-100.60	27.52	270
22	RAMOS ARIZPE	COAH	05032	-100.98	25.53	1399
23	MONCLOVA	COAH	05047	-101.42	26.90	645
24	SALTILLO	COAH	05048	-101.00	25.42	1520
25	MANZANILLO	COL	06018	-104.32	19.05	3
26	COLIMA	COL	06040	-103.73	19.23	495
27	COMITAN	CHIAP	07025	-92.13	16.25	1530
28	MOTOZINTLA	CHIAP	07119	-92.25	15.37	1455
29	CIUDAD DELICIAS	CHIH	08044	-105.43	28.20	1170
30	HIDALGO DEL PARRAL	CHIH	08078	-105.67	26.93	1950
31	CHINIPAS	CHIH	08167	-108.53	27.40	700
32	CAÑON FERNANDEZ	DUR	10004	-103.75	25.28	1300
33	LERDO	DUR	10009	-103.52	25.53	1135
34	CUENCAME	DUR	10012	-103.67	24.78	1580
35	EL SALTO	DUR	10025	-105.37	23.78	2538
36	FCO. I MADERO	DUR	10027	-104.30	24.47	1960
37	GUANACEVI	DUR	10029	-105.97	25.93	2200
38	NAZAS	DUR	10049	-104.12	25.23	1245
39	RODEO	DUR	10060	-104.53	25.18	1340
40	CELAYA	GTO	11009	-100.82	20.53	1754
41	DOLORES HIDALGO	GTO	11017	-100.93	21.15	1920
42	IRAPUATO	GTO	11028	-101.35	20.68	1725
43	OCAMPO	GTO	11050	-101.48	21.65	2250
44	SALVATIERRA	GTO	11060	-100.87	20.22	1760
45	SAN DIEGO DE LA UNION	GTO	11064	-100.87	21.47	2080
46	SAN JOSE ITURBIDE	GTO	11066	-100.40	21.00	2100
47	SANTA MARíA YURIRíA	GTO	11071	-101.15	20.22	1751

	STATION NAME	STATE	SMN ID	LONGITUDE	LATITUDE*	ALTITUDE+
48	PRESA VILLA VICTORIA	GTO	11082	-100.22	21.22	1740
49	SAN MIGUEL DE ALLENDE	GTO	11093	-100.75	20.92	1900
50	GUANAJUATO	GTO	11094	-101.25	21.02	2037
51	LEON (LA CALZADA, DGE)	GTO	11095	-101.68	21.08	1809
52	AYUTLA (CFE)	GRO	12012	-99.10	16.95	
53	IGUALA	GRO	12116	-99.53	18.35	635
54	HUICHAPAN	HGO	13012	-99.65	20.38	1102
55	SANTIAGO TULANTEPEC	HGO	13031	-98.37	20.08	2180
56	PACHUCA	HGO	13056	-98.73	20.12	2435
57	PRESA REQUENA	HGO	13084	-99.32	19.97	2109
58	ATEQUIZA (CHAPALA)	JAL	14016	-103.13	20.40	1520
59	CHAPALA	JAL	14040	-103.20	20.30	1523
60	EL FUERTE, OCOTLáN	JAL	14047	-102.77	20.30	1527
61	GUADALAJARA	JAL	14066	-103.42	20.72	1583
62	MAZAMITLA	JAL	14099	-103.02	19.92	2800
63	TEPALPA	JAL	14142	-103.77	19.95	2060
64	CD. GUZMAN	JAL	14500	-103.47	19.70	1535
65	APATZINGAN	MICH	16007	-102.35	19.08	682
66	PRESA COINTZIO	MICH	16022	-101.27	19.62	1997
67	CUITZEO DEL PORVENIR	MICH	16027	-101.15	19.97	1831
68	HUINGO	MICH	16052	-100.83	19.92	1832
69	JESUS DEL MONTE (MORELIA)	MICH	16055	-101.15	19.65	1950
70	LA CAIMANERA	MICH	16059	-100.90	18.47	287
71	PRESA LA VILLITA	MICH	16070	-102.18	18.05	
72	MORELIA (DGE)	MICH	16081	-101.18	19.70	1941
73	YURECUARO	MICH	16141	-102.28	20.35	1537
74	ZAMORA	MICH	16144	-102.28	20.00	1540
75	ZINAPECUARO	MICH	16145	-100.82	19.87	1840
76	ARTEAGA	MICH	16151	-102.28	18.35	860
77	CIUDAD HIDALGO	MICH	16152	-100.57	19.70	2000
78	URUAPAN	MICH	16164	-102.07	19.42	1610
79	ZACAPU	MICH	16171	-101.78	19.82	1986
80	ATLATLAHUACáN	MOR	17001	-98.90	18.93	1630
81	CUERNAVACA	MOR	17004	-99.25	18.92	1529
82	CUAUTLA	MOR	17005	-98.95	18.82	1291
83	PRESA EL RODEO	MOR	17006	-99.32	18.78	1100
84	ACAPONETA	NAY	18001	-105.37	22.50	22
85	AHUACATLAN	NAY	18002	-104.48	21.05	990
86	IXTLAN DEL RIO	NAY	18016	-104.37	21.03	1035
87	LAS GAVIOTAS	NAY	18021	-105.15	20.88	43
88	TEPIC	NAY	18038	-104.88	21.50	920
89	ALLENDE	NL	19003	-100.03	25.28	457
90	CERRALVO	NL	19010	-99.62	26.08	345
91	EL CUCHILLO	NL	19016	-99.25	25.73	145
92	HIGUERAS	NL	19025	-100.02	25.95	
93	ITURBIDE	NL	19027	-99.92	24.73	1480
94	LAMPAZOS	NL	19028	-100.52	27.03	320

	STATION NAME	STATE	SMN ID	LONGITUDE	LATITUDE*	ALTITUDE+
95	LOS RAMONES	NL	19042	-99.63	25.70	210
96	MIMBRES, GALEANA	NL	19047	-100.25	24.97	
97	MONTEMORELOS	NL	19048	-99.83	25.20	425
98	MONTERREY	NL	19052	-100.30	25.68	540
99	HUAJUAPAN DE LEON	OAX	20035	-97.78	17.80	1650
100	SANTA MARíA JACATEPEC	OAX	20042	-96.20	17.85	
101	JUCHITAN	OAX	20048	-95.03	16.43	46
102	MATIAS ROMERO	OAX	20068	-95.03	16.88	201
103	OAXACA DE JUAREZ	OAX	20079	-96.72	17.03	1550
104	SANTO DOMINGO TEHUANTEPEC	OAX	20149	-95.23	16.33	95
105	PIAXTLA	PUE	21063	-98.25	18.20	1155
106	PUEBLA	PUE	21065	-98.18	19.03	2209
107	TEZIUTLAN	PUE	21091	-97.35	19.82	2050
108	ZOQUITLAN	PUE	21114	-97.02	18.35	2140
109	PRESA CENTENARIO	QRO	22025	-99.90	20.52	1880
110	ALVARO OBREGON	QROO	23001	-88.62	18.30	
111	CHETUMAL	QROO	23032	-88.30	18.50	6
112	BALLESMI	SLP	24005	-98.93	21.75	30
113	CERRITOS	SLP	24008	-100.28	22.43	1150
114	CHARCAS	SLP	24010	-101.12	23.13	2020
115	MATEHUALA	SLP	24040	-100.63	23.65	1575
116	MEXQUITIC	SLP	24042	-101.12	22.27	2030
117	SAN LUIS POTOSI (DGE)	SLP	24069	-100.97	22.15	1870
118	TANZABACA	SLP	24090	-99.22	21.67	120
119	BOCATOMA SUFRAGIO	SIN	25009	-108.78	26.08	152
120	CULIACAN	SIN	25015	-107.40	24.82	62
121	CHOIX (DGE)	SIN	25019	-108.33	26.73	350
122	EL FUERTE	SIN	25023	-108.62	26.42	84
123	GUAMUCHIL	SIN	25037	-108.08	25.47	45
124	BADIRAGUATO	SIN	25110	-107.55	25.37	230
125	MAZATLAN	SIN	25135	-106.38	23.22	3
126	CIUDAD OBREGON	SON	26018	-109.97	27.50	35
127	PRESA LA ANGOSTURA	SON	26069	-109.37	30.43	50
128	TRES HERMANOS	SON	26102	-109.20	27.20	100
129	YECORA	SON	26109	-108.95	28.37	
130	HERMOSILLO	SON	26138	-110.97	29.07	200
131	(PRESA) PLUTARCO ELIAS CALLES	SON	26191	-110.63	29.93	
132	COLMACALCO	TAB	27009	-93.22	18.27	10
133	TAPIJULAPA	TAB	27042	-92.77	17.45	60
134	TEAPA	TAB	27044	-92.95	17.55	72
135	ABASOLO	TAM	28001	-98.37	24.05	61
136	MANTE (CAMPO EXPERIMENTAL INGENIO)	TAM	28012	-98.98	22.73	100
137	ANTIGUO MORELOS (EL REFUGIO)	TAM	28032	-99.08	22.55	242
138	MIGUEL HIDALGO	TAM	28038	-99.43	24.25	
139	MAGISCATZIN	TAM	28058	-98.70	22.80	90
140	SAN FERNANDO	TAM	28086	-98.15	24.85	43
141	TAMPICO (DGE)	TAM	28111	-97.87	22.22	12
142	VILLAGRAN	TAM	28118	-99.48	24.48	380

	STATION NAME	STATE	SMN ID	LONGITUDE	LATITUDE*	ALTITUDE+
143	SOTO LA MARINA	TAM	28152	-98.20	23.77	25
144	APIZACO	TLX	29002	-98.13	19.42	2404
145	TLAXCALA	TLX	29030	-98.23	19.32	2552
146	TLAXCO	TLX	29032	-98.13	19.63	2444
147	ANGEL R. CABADAS	VER	30011	-95.45	18.60	19
148	ATZALAN	VER	30012	-97.25	19.80	1842
149	CATEMACO	VER	30022	-95.10	18.42	338
150	CD. ALEMáN	VER	30025	-96.08	18.18	29
151	CHICONTEPEC	VER	30041	-98.17	20.98	595
152	IXHUATLAN	VER	30072	-98.00	20.70	306
153	JALTIPAN	VER	30077	-94.43	17.97	46
154	PAPANTLA	VER	30125	-97.32	20.45	298
155	RINCONADA	VER	30141	-96.55	19.35	313
156	VERACRUZ	VER	30192	-96.13	19.20	16
157	LAS VIGAS	VER	30211	-97.10	19.65	37
158	JALAPA	VER	30228	-96.92	19.53	1999
159	TUXPAN	VER	30229	-97.40	20.95	4
160	PANUCO	VER	30285	-98.17	22.05	60
161	PROGRESO	YUC	31023	-89.65	21.28	8
162	SOTUTA	YUC	31030	-89.02	20.60	11
163	MERIDA (DGE)	YUC	31044	-89.62	20.98	9
164	EL SAUZ	ZAC	32018	-103.23	23.18	2100
165	SOMBRERETE	ZAC	32054	-103.63	23.63	2300
166	JUCHIPILA	ZAC	32067	-103.13	21.42	1240
167	TEUL DE GLEZ. ORTEGA	ZAC	32070	-103.47	21.47	1900
168	ZACATECAS	ZAC	32086	-102.57	22.77	2450
169	ABILENE	TX	ABITX	-99.70	32.40	
170	EL PASO	TX	ELPTX	-106.50	31.80	
171	ELEPHANT BUTTE DAM	NM	EPBNM	-107.18	33.15	
172	PHOENIX	AZ	PHXAZ	-112.00	33.50	
173	SAN DIEGO	CA	SANCA	-117.20	32.70	
174	SAN ANTONIO	TX	SATTX	-98.47	29.53	
175	TUCSON	AZ	TUSAZ	-110.95	32.23	

Table 3.2. Spatially incomplete network of daily data stations for precipitation. The period of records for all the stations is from 1931 to 2001. * meters above sea level.

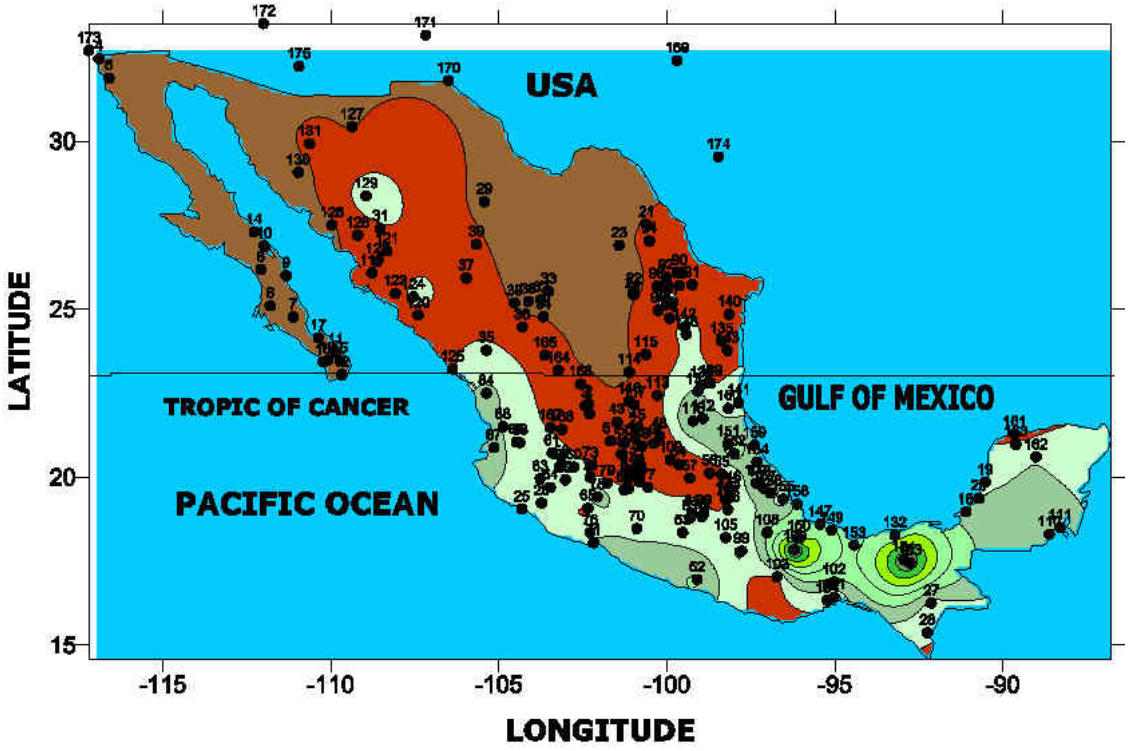


Fig. 3.5. Meteorological network of 175 with monthly precipitation data from 1931 to 2001 as used in the analysis of Principal Components (PC).

3.2.3 ENSO INDICES.

The Southern Oscillation Index (SOI).

One of the most typical measures utilised to explore the impacts of ENSO, is the Southern Oscillation Index (SOI). Since the 1800s this phenomenon had been observed as a difference in the sea-level pressures in the South Pacific, but its characteristics, extent and linked impacts in temperature and precipitation were not fully established by Walker and Bliss in the 1930s (Trenberth and Caron, 2000). Nowadays, it is widely accepted that the Southern Oscillation (SO) is a planetary-scale phenomenon, which involves an atmospheric mass of air in a standing wave shape, with a coherent exchange between the Eastern and Western hemispheres. The SO has its centre over Indonesia and the south tropical area of the Pacific Ocean. The SO is strongly associated with El Niño (EN), in this sense the cold phase is now called La Niña, while the warm phase is

frequently termed as El Niño, although their association is not always present. Nevertheless, the phenomenon is now universally referred as El Niño Southern Oscillation or ENSO (Ropelewski and Halpert, 1996).

The most extensively SO index recently used, because its correlation consistency, is the difference in sea level pressures between Tahiti and Darwin. In this research we are going to use the index defined by Ropelewski and Jones (1987). The index is calculated using five-month running means of the SOI that lie below the threshold of -0.5 standard deviations for more than five consecutive months; these cases considered "warm" episodes, and "cold" episodes are referred to the contrary conditions. Ropelewski and Jones (1987) state the post 1935 is a reliable source for ENSO related studies, and this condition makes it suitable with the purposes of the analysis.

Niño 3.4 Index

The high intensity of the ENSO events of the 1990s showed the necessity to extend the definitions of the four regions established in the 1980s. In this sense, Niño 3.4 (5° N - 5° S, 120°- 170° W) is today identified through Sea Surface Temperature (SST) anomalies centred approximately in the eastern half of the equatorial Pacific towards the west near the date line (fig. 3.6). up to date this index has proved to have the strongest link with ENSO-related impacts during the last decades (Barnston and Chelliah, 1997). Since April 1996 the measure also has allowed an improved scientific insight of the SSTs within the vital area between ENSO regions 3 and 4 (fig. 3.6). For the purposes of this research the standardised version of the Niño 3.4 index has been selected and extracted from the Climate Diagnostics Center (CDC) website: http://www.cdc.noaa.gov/ClimateIndices/.

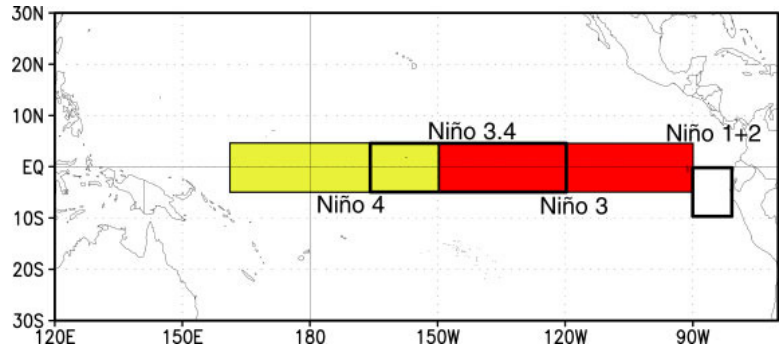


Fig 3.6. Current defined ENSO regions extracted from the Climate Diagnostics Center (CDC) website: http://www.cdc.noaa.gov/ClimateIndices/.

MULTIVARIATE ENSO INDEX (MEI)

Another option to explore the ENSO influence in a broader way (in the Mexican climate change context) is the Multivariate ENSO Index (MEI). The MEI is a more complete climatic measure when compared with the other ENSO indices available. The ocean and atmospheric variations are better considered by it, while it is also less vulnerable to the infrequent data errors of the monthly updating process. The index is computed as a weighted average of six different variables over the tropical Pacific, these parameters are: sea-level pressure (P), zonal (U) and meridional (V) surface winds, sea surface temperature (S), surface air temperature (A), and total cloudiness fraction of the sky (C). The MEI values are calculated for twelve sliding bi-monthly seasons (Dec/Jan, Jan/Feb, ... , Nov/Dec) based on the first unrotated Principal Component of the six combined fields of observation using the covariance matrix for the extraction, then standardised with respect to each season and considering 1950-93 as the reference period. More details about the index calculations can be found in Wolter (1987) and Wolter and Timlin (1993). Positive MEI values are linked to warm ENSO periods (El Niño), while negative values to cold periods of ENSO (La Niña). As this index is said to perform better at large-scale correlations (http://www.cdc.noaa.gov/people/klaus.wolter/MEI/table.html) and not necessarily at regional scales, It is expected that MEI can reflect better the relationships of the ENSO phenomenon with the meteorological variables chosen for this study, despite MEI incorporates more ocean and atmospheric parameters than the other indices,. In any case, MEI has been selected to check consistency in the results with those of the SOI and El Niño 3.4 indices.

3.3. MATHEMATICAL AND STATISTICAL METHODS APPLIED.

3.3.1. CONSIDERING DATA HOMOGENEITY.

It is well documented that small spatial and temporal variations or observational practices such as a slight change in the elevation of the station or the type of instrument could affect the consistency of records of a meteorological variable (Easterling et al., 1999). These changes could be reflected in the short or long term variation of the time series, and consequently influence the analysis of climate extremes variability, and their influence on the results can be significant, for Principal Component Analysis (see section 3.3.2) as well. For this reason, it is desirable to test the homogeneity of the stations selected before applying any analysis.

A time series is said to be homogeneous if all its fluctuations are caused by natural variability. In this sense, when an inhomogeneous time series is adjusted we are reducing the uncertainties of the results, and improving our understanding of the climate accordingly. The necessity of a precise scientific knowledge in this topic has recently increased its importance within the context of the study of climate change. Therefore, in applying the process of homogenisation to the data, utilising different techniques, we are searching for factors other than climate and weather. Although there is no single best technique, the approaches currently recommended to homogenise a time series are discussed in the following four steps (Aguilar et al., 2003):

1) Metadata analysis and quality control.
2) Creation of a reference time series.
3) Breakpoint detection.
4) Data adjustment.

For the analysis of homogeneity a detailed documentation of the history of the station is desired. For meteorological purposes the information about the data is called metadata. Knowledge of the station's history plays an essential part when preparing a high-quality dataset. Consequently, the reliability of the results is increased when the documentation

for the stations is available.

Metadata can help to identify changes in the conditions of the station. Among the changes that can be mentioned are: relocation, replacement of the instrument, exposure modifications, and changes in the recording procedures. Greater or lesser, all of them have a direct impact on the parameter values of the station. That is why a complete history of the station relates actual changes in the station with (gradual or sudden) observed changing patterns in the time series.

For the present study only digital instrumental data were used, in such a way that the objective was to extract the largest number of stations. Having this sort of digital information the available metadata was restricted to the most basic characteristics like station identifier, location, elevation and climatological normals. Other sources of metadata like changes in location, instruments, and observational practices were inaccessible to this research, making it extremely difficult to determine the artificial nature of some of the identified inhomogeneities.

Data quality control was addressed in section 3.2 of this chapter, as part of the process of detection of inhomogeneities. Daily comparisons, among the different digital databases were applied in order to find inconsistencies.

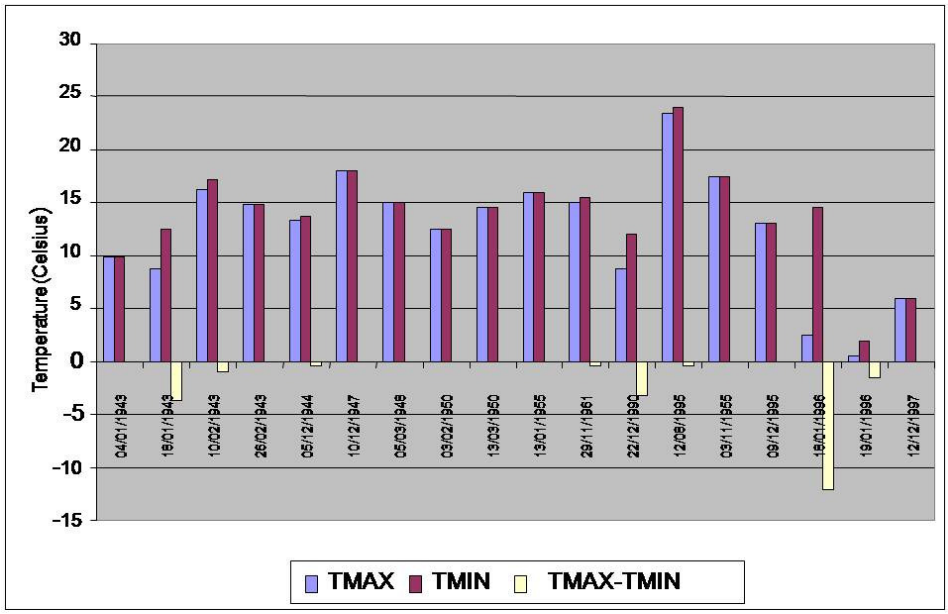


Fig. 3.7. Station with daily temperature errors before being corrected. In this case Tmin values are greater than Tmax.

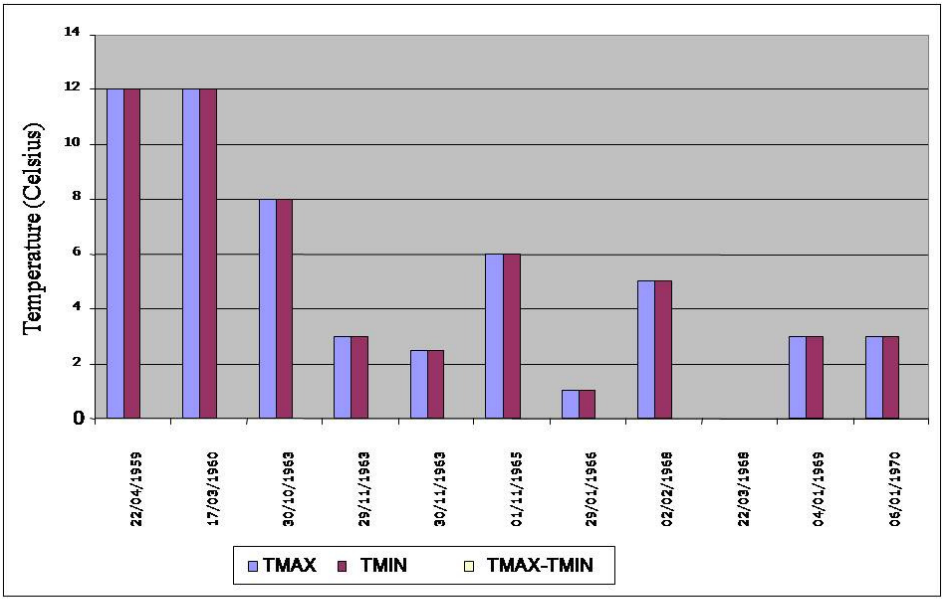


Fig. 3.8. Station with daily temperature errors before being corrected. In this case Tmin and Tmax have the same values.

Once the time series was ready, a set of basic statistics were computed like: mean, standard deviation, maximum and minimum to compare with other climatic studies in Mexico; these statistics were also used to easily identify outliers. Finally, annual precipitation, mean temperature and double-mass plots were prepared in this quality control process for every single time series to spot sudden changes in the climatic patterns. The Double-Mass plot is a technique utilised to find inconsistencies in a climatogical time-series. The underlying assumption is that the plotting of the accumulation of one quantity (a meteorological parameter at one station) against another during the same period will produce a straight line (45° slope) as far as the data are proportional. So, when a break is found that means a change in the constant of proportionality, or that the constant of proportionality is not the same at all rates of accumulation. Double-mass plots can be used to identify one or more inhomogeneities, and to correct them if the errors are clear enough (Cluis, 1983). An example of a plot after the application of this technique is seen in figures 3.9 and 3.10.

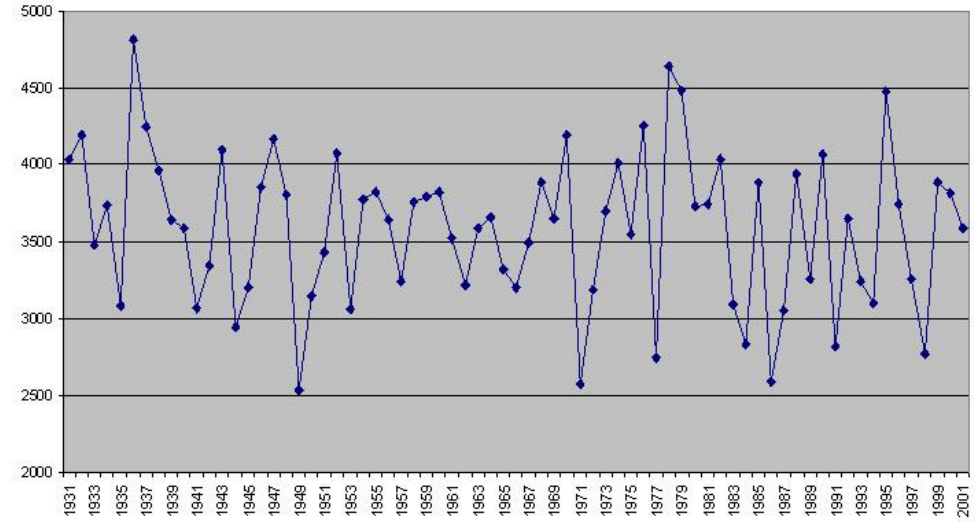


Fig. 3.9. Annual total precipitation (in mm) for station 27042; Tapijulapa, Tabasco.

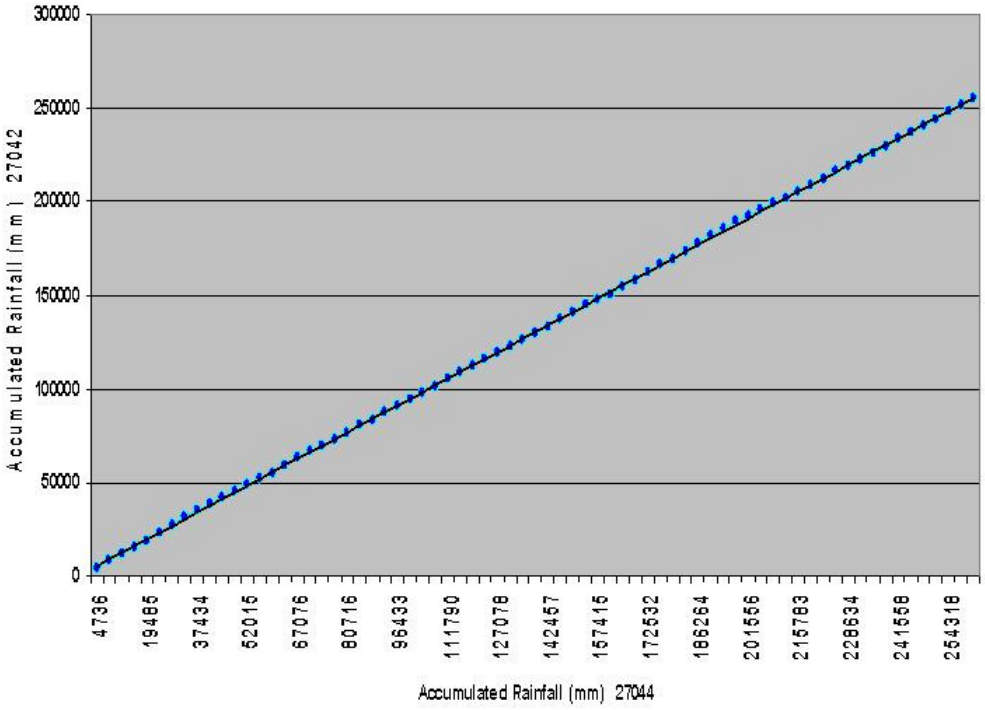


Fig. 3.10. Double Mass Plot for the station 27042; Tapijulapa, Tabasco.

No sudden jumps appear in Fig. 3.9 for station 27042 (Tapijulapa, Tabasco) after the quality control process described in section 3.2. After this analysis and "filtering out" evident errors like mistyped values, no major changes seem to have occurred in this location. The nearly "perfect" slope of the double mass plot of fig. 3.10 shows that the time-series of 27042 (against the data for station 27044) can be considered as a reliable source for the climatic analyses to be applied.

Another stage of *quality control of the data* was performed using the interactive program called RClimdex as an initial step to the extremes indices calculation. The main objective here was to identify possible mistyped errors that could affect the analysis. For instance, all precipitation values lower than 0 were considered as missing data; the same treatment was applied to the case in which daily minimum temperature was greater or equal than daily maximum temperature. Fig. 3.7 shows examples in which Tmin are equal or exceed the values of Tmax. Meanwhile the fig. 3.8 show examples in which Tmin values systematically are equal than Tmax, both set of data errors were corrected before applying subsequent analyses. The software is also able to identify outliers for a user-defined threshold, for the values of temperature (daily maximum and minimum temperature) the lower limit was set to the mean minus three standard deviations (mean - 3σ) and the mean plus three standard deviations (mean + 3σ) as the upper limit. All values beyond these thresholds were marked as suspicious and checked, then corrected accordingly when undisputable errors were present.

Due to the inherent characteristics of inhomogeneities -sometimes their variations are equal or even smaller than real natural climatic fluctuations- the process of detection is frequently difficult. To overcome this complexity it is recommended to create a reference time series. The most frequent way to construct them is to compute a weighted average using data from neighbouring stations or to select a section of surrounding stations whose data are considered homogeneous.

A clear regionalisation of the rainfall stations network made using PCA (see chapter 4) has facilitated the analysis of homogenisation. Having a group of stations that coherently varied across time made the comparisons easier. A weighted regional and individual time-series were prepared using the approach proposed by Jones and Hulme (1996). Using different indices like the Percentage Anomaly Index (PAI) and Standardised Anomaly Index (SAI) all the stations were plotted (See one example in fig. 3.11) searching for inhomogeneities.

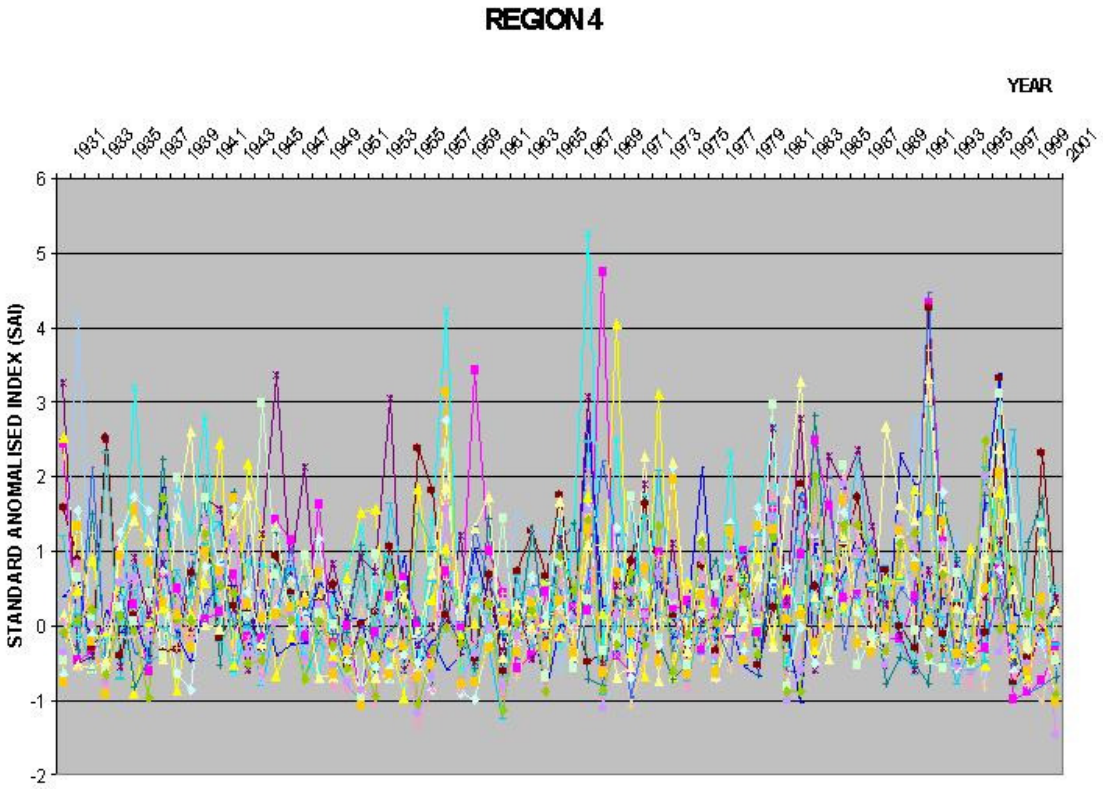


Fig. 3.11. Standard Anomalised Index (SAI) for the annual precipitation of all the stations of the resulting Region 4 after the Principal Component Analysis (PCA, see section 4.1).

In the process of calculation of the regional PAIs or SAIs, similarities include the possibility that the indices of the regions can generally avoid local effects. They share the same order of magnitude, are also designed to smooth sudden jumps in the series, and can identify the quasi-periodicity or modulation effect of large-atmospheric controls as can be fully observed in the very wet years of the late 1950s or the prolonged droughts of the 1990s. Among several differences, regional indices can preserve particularities inherent only to some regions like those along both coasts that are strongly impacted by hurricanes (as is the case of the north-eastern region hit by Hurricane Gilbert on 1988) or some areas by ENSO (like the north-western part of Mexico during the strong El Niño of 1982-83). Fig. 3.12 shows the calculated SAIs for the eleven regions extracted (of total annual precipitation) using Principal Component Analysis (see section 4.2).

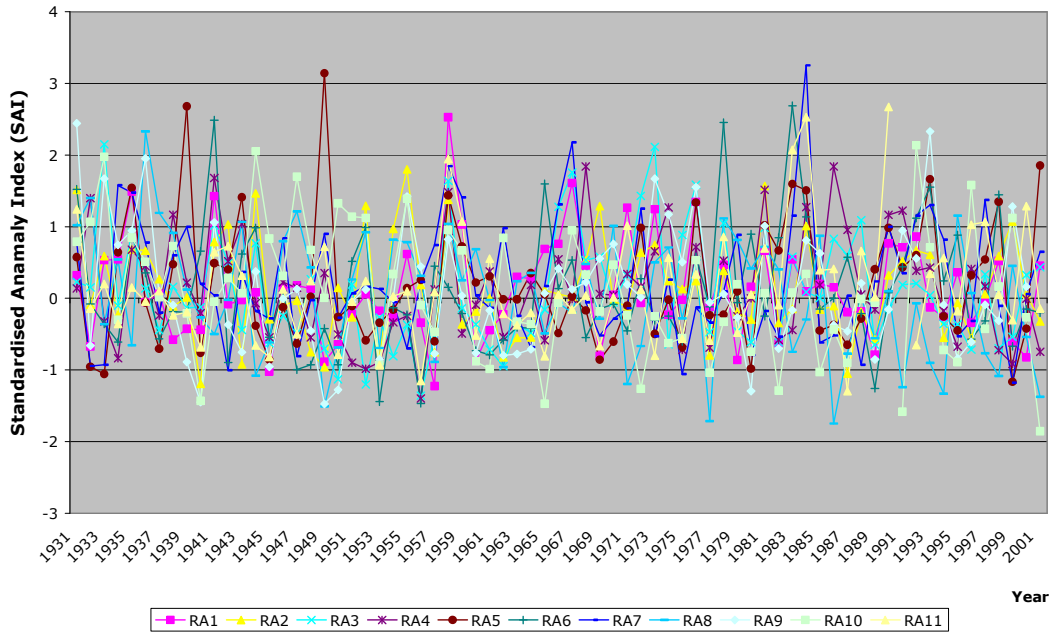


Fig. 3.12. Standard Anomalised Index (SAI) for the different regions (with total annual precipitation) after the Principal Component Analysis (PCA, see table 4.1).

Unfortunately, the detection of these inconsistencies for temperature using reference time-series was not feasible, as no clear results, i.e., coherent regions, were obtained with PCA (see section 4.3). So the construction of the weighted regional average was impossible. Another reason that impeded the comparisons among the stations for temperature was the sparsity of the network; neighbouring stations were not available for comparison of the dubious time-series. Finally, few homogeneous neighbouring temperature stations were ready to be used in this process.

Other indirect methods have been explored to identify undocumented inhomogeneities. If, it is not possible to build a reference time series, for reasons such as the sparsity of the network, there are alternative methods to identify the sorts of inhomegeneities within the data. In order to identify a sudden jump in a time series, common statistical methods like t-test are able to deal with the problem very well. If gradual artificial trends are involved like those caused by urbanisation, then regression analysis can perform better. For this study, the R-based program called RHtest was used to identify breakpoints. The approach of the program is the one outlined in Wang (2003). The objective of the two-phase regression model is to find a sudden changepoint (c) in the time-series. This undocumented breakpoint is found when:

$$F_{max} = \max_{1 \le c \le n} F_c$$

in which the changepoint c maximises F_c. For multiple changepoints c $\in$ {2,..., n-1} the F_c is computed as:

$$F_c = \frac{(SSE_{red} - SSE_{Full})}{SSE_{Full}/(n-3)}$$

under the null hypothesis of no changepoints and Gaussian errors ϵ_t, SSE_{Full} (the "full model" sum of squared errors) and SSE_{Red} (the "reduced model" of squared errors) are

defined as:

$$SSE_{Full} = \sum_{t=1}^{c}(X_t - \hat{\mu}_1 - \hat{\alpha}t)^2 + \sum_{t=c+1}^{n}(X_t - \hat{\mu}_2 - \hat{\alpha}t)^2$$

$$SSE_{Red} = \sum_{t=1}^{n}(X_t - \hat{\mu}_{Red} - \hat{\alpha}_{Red}t)^2$$

For this technique the case of a two-phase regression model with a common trend α ($\alpha = \alpha_1 = \alpha_2$) is considered, so the time-series is defined as:

$$X_t = \begin{cases} \mu_1 + \alpha t + \varepsilon_t, & 1 \le t \le c \\ \mu_2 + \alpha t + \varepsilon_t, & c < t \le n \end{cases}$$

In the context of climatology, extremes are singular events, within the limits of the dataset distributions having special weather conditions associated, that makes them of high interest for climatic studies. In order to assess these weather extremes daily data are essential. Until today there are only a few methods to correct sub-monthly inhomogeneities, Aguilar et al. (2003) give a good account of these techniques, although no recommendations are made to deal with extremes at these scales. Nevertheless, as it has been addressed in this section, several processes have been applied to identify the most obvious inconsistencies in the data, in order to avoid misleading results.

Finally, rapid urban growth is a possible factor for the increasing trend in temperatures across the globe. If we take the definition of urban as those places with a population greater than 50,000 (Easterling et al., 1997), we have that 8 stations for precipitation and 9 for temperature in Mexico fall under this condition. The urban heat island has been explored locally in tropical cities particularly in Mexico City by Jauregui (1995), or at regional and subregional scales by Englehart and Douglas (2003). Several procedures have been suggested by Karl et al. (1988) to correct this urbanisation temperature bias.

But when compared with the global average rise in mean temperatures, heat urban biases are relatively small (Karl et al., 1991). However, with the geographically widespread and accumulating evidence towards warming in temperatures, it is unlikely that urbanisation plays a key role in the upward trend (Karl et al., 1993). Principally because the SST average of the world is warming at a similar rate to the land average (IPCC, 2007). Urbanisation influences cannot be ignored at local scales, and care will be taken when evaluating the results on climate extreme indices for stations within urban areas.

Recent social and economic impacts of extreme events have highlighted the necessity of having more than a global network of average monthly climatic conditions. Extraordinary weather events require by definition long-term, and high-quality daily data. Although there are a few attempts to have a global set of daily data (Alexander et al., 2006; Vose et al., 2005; Easterling et al., 1999), there is a lack of a worldwide dataset that impedes the evaluation of climatic changes during the twentieth century (Karl and Easterling, 1999; Jones et al., 1999). This data deficiency is especially observed in tropical regions across the world (Easterling et al., 1997). Until the goal of a global database of daily data of the most important meteorological variables is reached, a set of widely accepted climatic extreme indices is being used instead (Alexander et al., 2006; Easterling et al., 2000). The development in this research of a set of Mexican climatological stations with spatial and temporal improved resolution permits the application of up-to-date methods to assess the secular behaviour of weather extremes in this country. This evaluation will contribute to a better understanding and comparison of the past climatic conditions, in a region encompassing tropical to subtropical regions within the context of a global changing climate.

3.3.2 PRINCIPAL COMPONENT ANALYSIS (PCA).

No matter what statistics and climatological normals could show us, non-linear behaviour and multi-dimensionality are still intrinsic, and even more important, frequently dominate the climate (Hannachi, 2004). In this context of complexity, how to extract the most important information behind a large set of meteorological stations with time discrete observations, and then make the data simpler to describe, is one of the basic questions within atmospheric sciences, and particularly in climatology. Principal Component Analysis (PCA) is the main technique to reduce the dimensionality.

Principal Component Analysis is a powerful multivariate analysis tool that reduces the high dimensionality of a dataset preserving as much as possible of the original variability of the data. In order to achieve these purposes PCA transforms the original set of observations to a new smaller group of pairwise uncorrelated variables (Principal Components, PCs) capturing the largest parts of the total variance. In that sense, the first member of the group or First PC is able to extract the highest fraction of the data variance, then the second Principal Component can obtain from the remaining variance the second highest part of the variability, and so on (Fig. 4.2).

The first PC ($\boldsymbol{\alpha}_1'\mathbf{x}$) is a linear function of the elements x (for p variables) with the largest maximum variance, $\boldsymbol{\alpha}_1$ is a vector of constants $\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p}$, and ′ meaning transpose (Joliffe, 2002), so the formula could be expressed as:

$$\boldsymbol{\alpha}_1'\mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j$$

In the same manner, k uncorrelated PCs ($\boldsymbol{\alpha}_1'\mathbf{x}, \boldsymbol{\alpha}_2'\mathbf{x}, \ldots, \boldsymbol{\alpha}_k'\mathbf{x}$) with the maximum variances in descending order can be extracted. A relatively small number (m<<p) of PCs containing most of the variance of the data is generally the result.

How are these PCs developed? Let Σ be the covariance (or correlation) matrix of the vector of random variables (or S for the variance of a sample), for each k=1, 2, …, p. The

k^{th} PC is defined by $z_k = \boldsymbol{\alpha}_k' \mathbf{x}$ in which $\boldsymbol{\alpha}_k$ is an eigenvector of Σ that corresponds to the k^{th} largest eigenvalue λ_k. If $\boldsymbol{\alpha}_k$ (sometimes called loading or coefficient) is conveniently chosen having unit length ($\boldsymbol{\alpha}_k' \boldsymbol{\alpha}_k = 1$, or normalisation constraint), then var $(z_k) = \lambda_k$ is the variance of z_k. The searching of the largest eigenvalue that maximises the variance of each k^{th} PC ($\boldsymbol{\alpha}_k' \mathbf{x}$) could then be expressed in general with the formula:

$$\text{Var}[\boldsymbol{\alpha}_k' \mathbf{x}] = \lambda_k \qquad \text{for } k=1,2,\ldots p.$$

In the early developments of PCA, unrotated techniques were the only option possible; this condition has gradually changed to the current wide spectrum of orthogonal and oblique rotated solutions which today allow better results to be produced. Unrotated solution techniques, as pointed by Richman (1986) are only suitable for application to those cases when weak simple structures are present and the PCs extracted have both positive and negative correlations throughout all the field of study. For this reason, although explored, unrotated techniques were explicitly disregarded in the present research as being useful for the final interpretations.

The resulting orthogonal PCs often allow easier interpretation than the original variables by reducing their dimensionality but conserving the highest possible variance, and therefore their most important characteristics. Indeed, simple structure is one of the most important characteristics of PCA. Its objective is to decrease the dimensions (p) of the original matrix in such a way than a linear composite of the m PCs found permits a concise scientific description of every variable (Richman, 1986).

It is precisely targeting simplicity in the physical interpretation that a technique for rotating PCs is used. Orthogonal solutions were first developed to overcome most of the unrotated techniques limitations, in particular VARIMAX has been extensively used in climatological studies; the special characteristic for orthogonal solutions in which each axis has to be normal to the rest has been frequently pointed out as artificial. In (rotated) orthogonal solutions like VARIMAX, QUARTIMAX and EQUAMAX the axes are

selected in such a manner that maximum variation along each axis is found, and also another condition is that any axis must be perpendicular to the others. Therefore, all these rotation methods try to define "important" components as those with the maximum absolute loadings, and are separated from the lowest ones. Loadings with moderate values (not easy for interpretation) are explicitly avoided.

Orthogonality is sometimes considered as a non-natural approach constraining the solution. Ignoring the orthogonal condition led to a new generation of techniques in which the restriction of perpendicularity was not present. For this reason, oblique (non-orthogonal) rotated solutions represented an alternative answer to unrotated and orthogonal solutions in PCA. Oblique methods like OBLIMIN or PROMAX try to define clusters and associate them precisely to only one component. This characteristic is frequently linked to the process of clarifying the interpretation when compared with orthogonal rotated solutions. In atmospheric sciences, oblique rotations are sometimes preferred to orthogonal solutions for their advantages in the interpretation of the results (Englehart and Douglas, 2002). DIRECT OBLIMIN has been frequently used amongst oblique rotations. Nevertheless, PROMAX permits clearer results in meteorology when a network with a large number of stations and high grade of complexity are found. So, one orthogonal (VARIMAX) and one oblique solution (PROMAX with kappa=2) were selected as suitable options to explore the complex climatic variability conditions of México.

It is known within PCA, and to be more specific in the simple structure rotation theory, that S-mode helps in regionalisation purposes. S-mode is only one of six different matrix configurations, in which the stations are the columns versus time that is the rows in the array.

In order to cope with contrasting climatic conditions in México: wet regimes in some south-eastern areas (total annual precipitation ≈ 4000 mm) and desert conditions in some regions of the north (total annual precipitation is sometimes less than 300 mm), the correlation instead of the covariance matrix has been used. Even, when we have variables

with the same units (mm) as for precipitation, large variance differences would dominate the low-order PCs; so the correlation matrices are preferred to covariance matrices for the PCA. Another reason to prefer correlation matrices is that covariance matrices are often chosen because of their easier interpretation for statistical inference, but given that the purpose of this regionalisation is purely descriptive as a preparation for further analyses, that advantage is not a factor for this study.

As this research has both an aim of regionalization of México but in contrasting climatic conditions, an obvious question arises: How many regions are sufficient to precisely describe Mexican climate? This discussion leads to the determination of the number of components to be retained.

Several studies have assessed the performance of single methods, or contrast the competence of a number of different techniques, but there is no consensus about the best method for determing the most significant number of principal components (Peres-Neto et al., 2005; Al-Kandari et al., 2005). Because of the size and complexities of the datasets, a PCA graphical tool called the Scree Test is used in this thesis. The component numbers are the abscissa in the plot and their corresponding eigenvalues the ordinates. The plot is seen as a mountain in which the slope is formed by the "true number" of factors containing most of the variance, and the foot by the random components. Therefore, the foot of the mountain or scree straightens closely matching a line at the end of the plot. The aim is to find the last evident break before the variance between components becomes negligible (Cattell, 1966). The low-order PCs before this point of inflexion are then considered as the most relevant and meaningful for the study.

The determination of the number of PCs and therefore of climatic regions in Mexico has also required a careful classification, i.e., to assign each one of the stations to only one of the resulting regions. To comply with this requirement a strict rule was set of only accepting absolute loadings greater than 0.4. (White et al., 1991). So, according to this, the largest value in the loadings (or primary pattern) matrix clearly defines its corresponding component and consequently the region to which the station belongs. With

the same classification purpose in mind, the 'eigenvalue one' criterion was applied (Mather, 1976), i.e. only eigenvalues greater than 1.0 were considered for the extraction. The reason behind this is that, when is normalised each variable has a intrinsic variance equal to unity, every eigenvalue less than one should then be discriminated, and not worthy to be considered in the analysis. Finally, recalling that the missing values total was restricted to less than ten per cent for every station in the network and replaced with the long-term mean, the election of pairwise or listwise deletion has no influence on the final results.

All methods of rotation overcome the disadvantages of unrotated solutions. Among the drawbacks of these non-rotated solutions we can list the following:

Geographically dependent results. It is a well known phenomenon that sometimes topography has a strong influence on the delineation of contours. For some meteorological variables like precipitation, altitude exerts a linear response. This characteristic is frequently observed in the loading patterns of the PCs across an area using unrotated solutions.

No stability. In order to prove the consistency of the results, sometimes the data are divided into subdomains of the original variables. For example, a group of stations could be classified geographically taking into account their coordinates, in which a latitudinal or longitudinal line could represent a boundary. Regardless of any subdivisions, PCA patterns should be in accordance to the results when the whole domain is considered (Comrie and Glenn, 1998).

Closed Eigenvalues. When extracted eigenvalues are so closely spaced, most of the time, unrotated methods are unable to precisely separate PCs. Even worse, sometimes this problem becomes so difficult that eigenvalues could be mixed among them.

Artificial Results. Unrotated solutions could produce patterns that don't have a physical basis, i.e. *Buell patterns*. This is particularly true when from a previous insight to the data

a well known configuration is expected. Richman and Lamb (1985) shows an example in which PCs two to 10 are not completely in accordance with the observed patterns before the analysis.

Regardless of orthogonal or oblique solution ease of interpretability classifies the degrees of simple structure as strong, moderate or weak. The amount of simple structure is best explored through pairwise plots of the resulting coefficients. In theory a strong simple structure unveils a hidden order in the data.

Among the applications of PCA that can be mentioned are:

- Identification of groups of variables that vary coherently in a dataset.
- Reduction of the original dimension of the dataset, resulting in a smaller and independent set.
- PCA is able to eliminate redundancy in the original variables.
- It could be considered as a preliminary step of cluster analysis. PCA clarifies the clustering by eliminating the eigenvectors with the lowest-valued eigenvalues.
- PCA is an alternative to the construction of a set of linear functions of the original variables; as opposed to a process based solely on *a priori* judgements.
- The possibility to spot a new group of individuals varying coherently, that other method cannot successfully achieve.
- Principal Component Analysis could help to easily identify “outliers”, i.e. individuals that are behaving clearly different to the other variables in a group.

- PCA could be considered as a preliminary tool to multiple regression analysis. The resulting components could be used as an approach of a set of regressor variables.

3.3.3. REGIONAL AVERAGES.

In performing PCA across the network, the objective was to find different groups of stations that are varying coherently across time. The amplitude of a particular PC will incorporate all the stations. Here we want to calculate a regional average, based on PCA, but just with the stations in a region. When calculating regional averages we want the dominant time-series features of the sites to remain. Also, we are trying to avoid local factors like topography. We use the approach suggested by Jones and Hulme (1996) to compute regional averages. Among the different indices proposed, the Standardised Anomaly Index (SAI) has been selected, to be consistent with the extracted ENSO indices (See section 3.2.3). Standardised anomalies are first calculated for each station as:

$$\Delta\hat{P}_{ik} = \frac{P_{ik} - \overline{P}i}{\sigma_i}$$

where $\Delta\hat{P}_{ik}$ is the standardised anomaly for year k at station i from a group of N stations, in accordance with the resulting regions of PCA (section 4.3.1). $\overline{P}i$ and σ_i are mean and standard deviation of the station i respectively (based on a common period which is the total length of the time-series).

The weighted regional SAI is computed as follows:

$$\left\langle \Delta\hat{P}_k \right\rangle = \sum_{i=1}^{N} w_i \Delta\hat{P}_{ik}$$

In this equation $\left\langle \Delta\hat{P}_k \right\rangle$ is the regional standardised anomaly for year (month) k. The

weights are obtained as the long-term ratio of the local ($\overline{P_i}$) to regional $\langle \overline{P} \rangle$ means:

$$w_i = \frac{\overline{P_i}}{\left\langle \overline{P} \right\rangle}$$

where the long-term mean (of the total N years) for a single station (i) is defined as:

$$\overline{P_i} = \frac{1}{N} \sum_{k=1}^{N} P_k$$

and the mean (of all the N stations) of a region as:

$$\left\langle \overline{P} \right\rangle = \frac{1}{N} \sum_{i=1}^{N} \overline{P_i}$$

In order to test the stability of the regional values a different weight was utilised

$$w_{ik} = \frac{P_{ik}}{P_{k(total)}}$$

in which

$$P_{k(total)} = \sum_{i=1}^{N} P_{ik}$$

is the sum of the precipitations of all the stations in a region, for a given year (month) k.

The two different results were compared year by year, finding very similar results. Therefore, the first approach was used in the subsequent analyses.

It is important to notice here that, for the regional averages the same seasonal definitions

established in section 2.2.1 were applied here, i.e. total annual precipitation, wet (May-Oct) and dry (Nov-Apr) seasons. These time series would also be used in our ENSO-related research. But monthly time-series are also available, especially for lag correlation analyses.

3.3.4. EXTREME WEATHER ANALYSIS.

During the 2005 hurricane season in Mexico, tropical cyclone Stan struck the south-eastern part of Chiapas State, and later Hurricane Wilma hit the Mexican Atlantic coast around the tourist city of Cancún. There was a perception with the public, influenced by the media that extraordinary events were occurring. The question for the scientific community, however, is: Are the intensity and frequency of extreme events increasing and if so is this related to anthropogenic influences on the climate system? To scientifically evaluate these sorts of climatic questions is very difficult. What is important first is to be sure that the climatic series are of good quality.

Average climatic conditions and their variability have been extensively explored recently; this is especially true in the case of anthropogenic climate change (Easterling et al., 1999). Mean conditions of the climate do not give a complete picture; they just tell us part of the history of the changing regional climates of the world. Other aspects of these meteorological parameters need to be explored if we are to understand the underlying processes of the climate system. Amongst the important characteristics that can be assessed are the weather extremes, because they are a good measure of the rapid change of climate, and also they generally have a great impact on society in general. Greater trends in extremes compared to the mean temperature trends were found in an analysis applied to long time series from Europe and China (Yan et al., 2002). Unfortunately, studies on climate extremes using daily data are still relatively scarce, but improvements and extension into unanalysed areas are gradually being made.

Time series of monthly data are sufficient to explain changes in the climatological normals and their variability on similar or longer time scales (Jones et al., 1999). These

databases are satisfactory for documenting the climatic history of the recent warming at hemispheric and global scales. But, as recent years have shown in different regions of the world, there appear to be more extremes occurring (Alexander et al., 2006). Nevertheless, unequivocal proofs of these fluctuations in weather extremes are necessary to support the accumulating evidence.

Even in developed countries with potential for large climatic databases like the USA and Canada, there is still a deficiency of homogeneous climatological time-series to evaluate the recent secular behaviour of the extremes in this region (Easterling et al., 1999). Ironically the analysis of extremes can also help to highlight that monthly-based homogeneity analyses are inadequate (Yan et al., 2002). Fewer studies exist dealing with extreme weather in developing countries. This situation is being rectified and a recent study by Alexander et al. (2006) analyses extensive datasets. This work stems from developed datasets in specific regions: Africa (New et al., 2006), South America (Haylock et al., 2006), South East Asia and the South Pacific (Manton et al., 2001), Central America and northern South America (Aguilar et al., 2005), and Central and South Asia (Klein Tank et al., 2006).

As a country, Mexico is not not well represented or definitively absent in the climatic extreme analyses. The very few assessments of the changing climate in the country were made as part of a global evaluation or the North American region (e.g. Alexander et al., 2006; Vose et al., 2005; Easterling et al., 1999). When dealing with climatological monthly data as well as for evaluating extremes, part of the problem is the geographical sparsity of the set of stations with suitable long-term time-series of daily data. In Mexico, the Servicio Meteorológico Nacional (Mexican Meteorological Service) maintain a network that has remained unchanged assuring relatively long records with minor variations (Easterling et al., 1999); but these data have only generally been kept in manuscript form. A key factor that contributed to the development in this field of science was the needs of the Intergovernmental Panel on Climate Change (IPCC) to monitor firstly the mean climatic state of the world and secondly to evaluate the trends in extreme weather at national, regional and global scales.

There still is not a single way to define an extreme in climate. Up to today climatologists continue dealing with the problem of isolating changes due to sampling, station location, and indisputable changes in extremes (Frich et al., 2002). For these reasons, several attempts have been made to build a scientific consensus in the analysis of weather extremes. Unfortunately, it is very frequent that these extraordinary events also have socio-economic impacts, deeply affecting the way they are perceived. Therefore, not only scientific but sometimes socio-economic considerations have played an important role in the process of defining climatic extremes.

The lack of climate extremes definitions has gradually been overcome. For studying weather extremes across the USA, Karl et al. (1996) defined an index which was termed the Climate Extremes Index based not only on the exceedence of thresholds for meteorological variables (such as temperature or precipitation), but also the percentage of the country affected by severe drought. Following on from this, Beniston and Stephenson (2004) developed a set of characteristics (not mutually exclusive) that can measure extremes. These are listed in there study as follows:

- how rare they are, which involves notions of frequency of occurrence;
- how intense they are, which involves notions of threshold exceedence; and
- the impacts they exert on environmental or economic sectors in terms of costs or damages.

They also point out the way in which weather extremes have been defined in the Third Assessment Report of the IPCC (2001) in terms of frequency, as several meteorological variables (precipitation, wind velocity or temperature) exceed the 10% or 90% quantiles of their distribution. But it really was when the IPCC 2nd Assessment report identified the deficiency of studies on trends of daily data and climate extremes that these efforts significantly increased in scale: locally, regionally and globally (Alexander et al., 2006; New et al., 2006; Haylock et al., 2006). Since then a group of climatologists, The Expert Team (ET) on Climate Change Detection and Indices (ETCCDI) have been conducting an international effort to develop, calculate and analyse a set of indices to standardise and compare the results globally (http://cccma.seos.uvic.ca/ETCCDMI/index.shtml). Data

For Precipitation

PRCPTOT	**Wet-day precipitation**	**Annual total precipitation from wet days**	**mm**
SDII	Simple daily intensity index	Average precipitation on wet days	mm/day
CDD	Consecutive dry days	Maximum number of consecutive dry days	days
CWD	Consecutive wet days	Maximum number of consecutive wet days	days
R10mm	Heavy precipitation days	Annual count of days when RR>=10mm	days
R20mm	Very heavy precipitation days	Annual count of days when RR>=20mm	days
R95p	Very day wet precipitation	Annual total precipitation when RR>=95th percentile of 1961-1990	mm
R99p	Extremely wet day precipitation	Annual total precipitation when RR>=99th percentile of 1961-1990	mm
RX1day	Max 1-day precipitation	Annual maximum 1-day precipitation	mm
RX5day	Max 5-day precipitation	Annual maximum 5-day precipitation	mm

For Temperature

FD	**Frost days**	**Annual count when TN(daily minimum)<0˚C**	**days**
SU	Hot days	Annual count when TX(daily maximum)>25˚C	days
ID	Cold days	Annual count when TX(daily maximum)< 0˚C	days
TR20	Warm nights	Annual count when TN(daily minimum)> 20˚C	days
GSL	Growing season length	Annual count between first span of at least 6 days with TG>5˚C after winter and first span after summer of 6 days with TG<5˚C	days
TXx	Hottest day	Monthly highest TX	˚C
TNx	Hottest night	Monthly highest TN	˚C
TXn	Coolest day	Monthly lowest TX	˚C
TNn	Coolest night	Monthly lowest TN	˚C
TN10p	Cool night frequency	Percentage of days when TN<10th percentile of 1961-1990	%
TX10p	Cool day frequency	Percentage of days when TX<10th percentile of 1961-1990	%
TN90p	Hot night frequency	Percentage of days when TN>90th percentile of 1961-1990	%
TX90p	Hot day frequency	Percentage of days when TX>90th percentile of 1961-1990	%
WSDI	Warm spell day index	Annual count of days with at least 6 consecutive days when TX>90th percentile of 1961-1990	days
CSDI	Cold spell day index	Annual count of days with at least 6 consecutive days when TN<10th percentile of 1961-1990	ysda
DTR	Diurnal temperature range	Monthly mean difference between TX and TN	˚C

Table. 3.3. Weather Extreme Indices as defined by the Expert Team (ET) on Climate Detection and Indices (ETCCDI) and tabulated in New et al. (2006).

quality and calculations can be performed using the free statistical package "R" (http://www.r-project.org) through a graphical-interfaced program called "RClimDex". The current core indices - as defined by the ET and tabulated in New et al. (2006) - are:

REFINING THE DATA SELECTION FOR EXTREME ANALYSIS

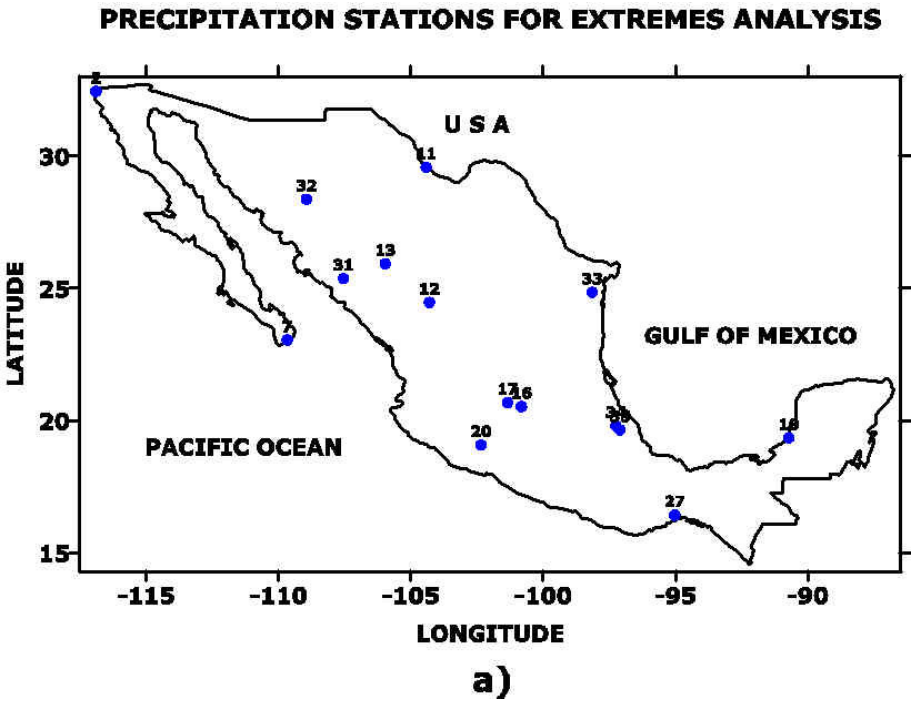
Although the selection of stations nearly replicates the process of the Data Extraction (see section 3.2.); a few additional characteristics needed to be introduced in order to comply with the slightly more particular conditions necessary for the analysis of weather extremes. As mentioned, meteorological daily records are practically indispensable in the analysis of extremes. *Originally, daily temporal resolution was targeted for the data extraction*. However, during the process of reviewing and choosing the suitable stations to be analysed many of them were incomplete with some missing data. These data were filled with their corresponding monthly averages of the same stations whenever it was available (see section 3.3.2). This means that only a relatively small number of time-series are free of unfilled data.

Daily data with low percentages of unfilled data were preferred when selecting the stations to calculate the extreme indices. Given that good spatial coverage was obtained for the Principal Component Analysis (PCA) (section 4.1) for the network of monthly precipitation, rainfall was used as the reference database for the determination of both: the best daily records of temperature and precipitation. In order to compare the extreme analysis with the PCA results, at least one station was desirable to be selected per (precipitation) region. A contrasting assessment could then be made between regional and local scales. The main objective is to obtain for daily data the same database as that set of monthly rainfall data used in the analysis of PC. A comparison will then be possible between the regional time series –constructed from the results of PCA- and the single station data, and hopefully find inconsistencies or differences between their climatic patterns. The resulting set of stations for both meteorological variables is listed in Table 3.2.

	station name		longitude° W	latitude° N	precip	temp	altitude*	pop+
1	PABELLON DE ARTEAGA	AGUASCALIENTES	-102.33	22.18		X	1920	34.296
2	PRESA RODRIGUEZ	BAJA CALIFORNIA	-116.9	32.45	X	X	100	1210.82
3	COMONDú	BAJA CALIFORNIA SUR	-111.85	26.08		X	260	63.864
4	EL PASO DE IRITU	BAJA CALIFORNIA SUR	-111.12	24.77		X	140	196.907
5	LA PURíSIMA	BAJA CALIFORNIA SUR	-112.08	26.18		X	95	11.812
6	SAN BARTOLO	BAJA CALIFORNIA SUR	-109.85	23.73		X	395	
7	SAN JOSE DEL CABO	BAJA CALIFORNIA SUR	-109.67	23.05	X		7	105.469
8	SANTA GERTRUDIS	BAJA CALIFORNIA SUR	-110.1	23.48		X	350	
9	SANTIAGO	BAJA CALIFORNIA SUR	-109.73	23.47		X	125	
10	CHAMPOTON	CAMPECHE	-90.72	19.35	X		2	70.554
11	OJINAGA	CHIHUAHUA	-104.42	29.57	X		841	24.307
12	FCO. I MADERO	DURANGO	-104.30	24.47	X		1960	
13	GUANACEVI	DURANGO	-105.97	25.93	X		2200	10.794
14	EL PALMITO	DURANGO	-104.78	25.52		X	1630	6.011
15	SANTIAGO PAPASQUIARO	DURANGO	-105.42	25.05		X	1740	43.517
16	CELAYA	GUANAJUATO	-100.82	20.53	X		1754	382.958
17	IRAPUATO	GUANAJUATO	-101.35	20.68	X	X	1725	440.134
18	PERICOS	GUANAJUATO	-101.1	20.52		X	1772	226.654
19	SALAMANCA	GUANAJUATO	-101.18	20.57		X	1722	226.654
20	APATZINGAN	MICHOACAN	-102.35	19.08	X		682	117.949
21	CUITZEO DEL PORVENIR	MICHOACAN	-101.15	19.97		X	1831	26.269
22	HUINGO	MICHOACAN	-100.83	19.92		X	1832	48.917
23	CIUDAD HIDALGO	MICHOACAN	-100.57	19.7		X	2000	106.421
24	ZACAPU	MICHOACAN	-101.78	19.82		X	1986	69.7
25	AHUACATLAN	NAYARIT	-104.48	21.05		X	990	15.371
26	LAMPAZOS	NUEVO LEON	-100.52	27.03		X	320	5.305
27	JUCHITAN	OAXACA	-95.03	16.43	X		46	325.295
28	MATIAS ROMERO	OAXACA	-95.03	16.88		X	201	75.095
29	SANTO DOMINGO TEHUANTEPEC	OAXACA	-95.23	16.33		X	95	217.624
30	MATEHUALA	SAN LUIS POTOSI	-100.63	23.65		X	1575	78.187
31	BADIRAGUATO	SINALOA	-107.55	25.37	X	X	230	37.757
32	YECORA	SONORA	-108.95	28.37	X		1500	6.069
33	SAN FERNANDO	TAMAULIPAS	-98.15	24.85	X	X	43	57.412
34	ATZALAN	VERACRUZ	-97.25	19.80	X	X	1842	48.179
35	LAS VIGAS	VERACRUZ	-97.10	19.65	X	X	37	14.161

Table 3.2. Daily data stations for temperature and precipitation for extreme analysis. The period of records for all the stations is from 1941 to 2001. * meters above sea level. + Population in thousands.

Thirty five stations were selected for the extreme analysis: 15 of those time-series have daily precipitation and 26 temperature data. Unfortunately for comparison purposes, only six climatological stations have good enough data for both meteorological variables. The period of the records for the analysis starts in 1941 and ends in 2001. The lengths of records for precipitation have been reduced for this study to begin in 1941 instead of 1931 as for the monthly records. The reason behind this decision is that there should be at least one climatic representative station containing daily data per PCA region (see chapter 4). This is true for all regions except those from region 7 to 11. Climatic regionalisation using PCA had clear results for annual rainfall; that is why, a time-series per PCA resulting region was computed utilising weighted averages, besides selecting one climatic representative station per region. This permits comparison between regional and local scales for rainfall. However, no clear results (no clear PCA regions) were obtained for temperature; this means no PCA regions could be used. For this reason, the 175 station network with monthly precipitation (see section 4.2.1.) was then considered as a reference for the extraction of the largest number of temperature stations. Despite the limitations, only some north and south-eastern areas of the country were not covered for the extreme analysis. The spatial coverage of both networks is displayed in figure 3.6 a) for precipitation, and 3.6 b) for temperature.



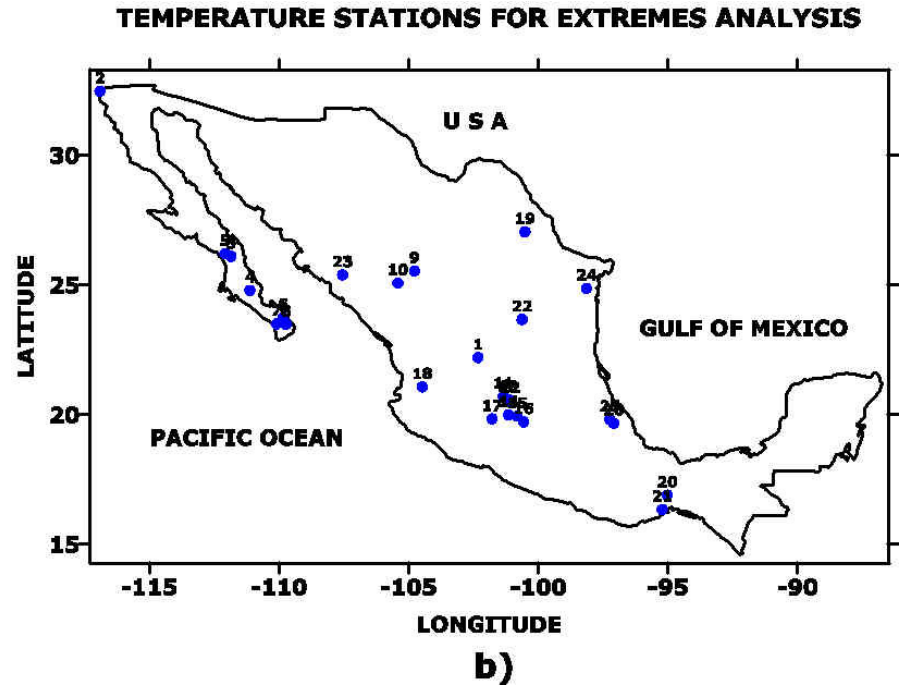


Fig 3.13. Network of a) precipitation and b) temperature stations with daily data for the analysis of extremes (in accordance with table 3.2). The period of the records is from 1941 to 2001.

Three main factors dominated the selection process of the time-series for the extreme analysis: daily data, the length of the records and the completeness (low numbers of missing values). However additional characteristics considered were the possible influence on extremes from: altitude, homogeneity and urbanisation. As discussed in section 4.2, even though the altitude effect is explicitly avoided (using the ratio of the precipitation of each station to its long-term mean) for the PCA on precipitation, high elevation could still exert its force in the atmospheric phenomena. It is interesting to note that 6 rainfall and 10 temperature stations exceed the 1000 m.a.s.l. threshold.

3.3.5 CORRELATION ANALYSES.

Non-parametric Correlations.

Frequently, the task of scientists is to establish relationships between two or more variables. A correlation measures the linear relationship between variables (Field, 2005). The most widely used method (for the complexity of their calculations, non-parametric were more complicated than linear correlations, it is not until recently that computers have overcome with this limitation) to evaluate linear correlations is the Pearson product-moment correlation coefficient. Although a linear correlation coefficient can often give an approximate idea of the strength of the relation between the variables under study, it has a limited resistance and robustness, and also lacks reliability in the determination of the level of significance (Haylock, 2005). Rank (or Non-parametric) correlation coefficients can overcome these limitations; normally distributed data is also not a condition for these techniques.

In contrast to the linear correlation, a non-parametric correlation coefficient measures association, i.e. a monotonic relationship between variables. Very well known measures of association are the Spearman rank-order and Kendall's tau correlation coefficients. Because Kendall's tau deals better (than Spearman's) with small datasets and a large number of tied ranks (Haylock, 2005), this is the non-parametric correlation coefficient that will be used to test the strength of the relationships and level of significance between two variables in this thesis.

Kendall's tau-b (τ) is a non-parametric correlation that measures the association of the number of concordant and discordant pairs of observations. A pair of values is said to be concordant if the vary together, and discordant if the vary differently. The coefficient ranges between -1 (ranks increasing separately) and +1 (ranks increasing together). The formula for Kendall's tau-b is:

$$\tau = \frac{\sum_{i<j} \operatorname{sgn}(x_i - x_j)\operatorname{sgn}(y_i - y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}}$$

where

$$T_0 = \frac{n(n-1)}{2}$$

$$T_1 = \sum \frac{t_i(t_i - 1)}{2}$$

$$T_2 = \sum \frac{u_i(u_i - 1)}{2}$$

t_i is the number of tied x values in the i_{th} group of tied x values, u_i is the number of tied y values in the j_{th} group of tied y values, n is the number of observations and sgn(z) is defined as:

$$\operatorname{sgn}(z) = \begin{cases} 1 \, if \; z > 0 \\ 0 \, if \; z = 0 \\ -1 \, if \; z < 0 \end{cases}$$

The main advantages of Kendall's tau-b are that the distribution has slightly better statistical properties and also being defined in terms of probability of concordant and discordant pairs of observation, this non-parametric correlation coefficient leads to a direct interpretation of the results (Chrichton, 2001).

Lag Correlation.

Responses in meteorological parameters (e.g. rainfall) to changes in large-scale phenomena are not immediate. It sometimes takes a time period the scale of months to seasons for an ocean-atmospheric process, like ENSO (El Niño Southern Oscillation), to be fully developed. Then, for its scientific understanding, it is crucial to address these delayed modulations when evaluating these atmospheric relationships.

The Cross-Correlation function (or lag cross-correlation) is a suitable technique to measure the time shifts between the continuum and lag variations (Chatfield, 1991). Its main purpose is to find the lag that maximises the coherence (linear correlation) between two time series. When this tool is only applied to the same variable is called AutoCorrelation Function. Lag Cross-Correlations must be compiled for all lags (positive or negative), in such a way that a significant maximum correlation is found for a specific time shift (τ_k). The formula that explains these relations between the variables is:

$$CCF(\tau_k) = \frac{(1/N)\sum_{i=1}^{N-k}(x_i - \bar{x})(y_{i+k} - \bar{y})}{\sqrt{(1/N)\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{(1/N)\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

where the lag τ_k is the size of the time shift: $\tau_k = k\Delta t$, k=0,1,…,N-1 and $\bar{x}$ y $\bar{y}$ are the means of y_i and x_i.

This definition implicitly assumes that both time-series are stationary in their means and variances in a sample of N pairs of values. A direct dependency expressed in the linear correlation between the variables is also expected; nevertheless, this relationship could be substituted by a non-parametric correlation (e.g. Spearman's rho). Although CCF is probably not the best estimator (Welsh, 1999), it is mainly utilised because of its efficiency and consistency.

CCF has among other limitations, the following:

- It is defined in terms of linear dependency. This restriction is really artificial; a non-linear approach could lead to better results.
- Because of its lack of robustness, non-parametric tests can be more useful than the linear correlation technique used in the formula.

3.4. CONCLUSIONS.

Studies on climate change using daily data are scarce in developing countries. The research in this area needs reliable information, especially long-term time-series. In Mexico there have been several efforts to develop a national digital database of climatological data. Unfortunately, those databases still lack sufficient geographical coverage and analysis of data quality. Therefore, in order to contribute to the understanding of the climatic patterns of Mexico within the context of global warming, it was necessary to construct a national high-quality database of rainfall and temperature at monthly and daily time-scales.

A network of 175 rainfall and 52 temperature stations with monthly data, with good spatial coverage has been prepared to study climate change patterns in Mexico. The meteorological time-series have been extracted from six different digital sources, and a process of inter-comparison has been applied among them. Monthly data for precipitation has 71 years of information from 1931 to 2001; meanwhile the length of the daily data (rainfall and temperature) series is ten years shorter, spanning 1941 to 2001. The maximum fraction of missing values was restricted to ten per cent. Climatically speaking, in most of the country the precipitation is concentrated during the months of May to October, this period was considered as the wet season, while the interval from November to April was called the dry season. These definitions and the computing of annual figures were also applied to temperature. In addition, basic statistical properties

like: mean, standard deviation, maximum and minimum values were calculated to assure the reliability of the results of the analyses in this research.

Three different ENSO-related indices of this phenomenon have been selected in order to test their relationships with the rainfall and temperature across the country. They are the Southern Oscillation Index (SOI), Niño 3.4 and the Multivariate El Niño index (MEI), all of them are expressed in a standardised form in order to avoid external influences. The decades of 1980s and 1990s have seen a period of increasing intensity and frequency of ENSOs, with accumulating evidence of global warming. Therefore, the extracted ENSO indices are expected to be strongly linked to the rainfall and temperature in Mexico at regional and local scales.

The meteorological variables extracted are expressed at both monthly and daily time scales. These temporal scales have had direct influence in the methods selected and applied for this research. For instance, PCA is able to unveil a hidden order among a set of variables. Climatically speaking, one of the most important properties is that this method can find groups of stations varying coherently. Technically, rotated are more efficient than the unrotated solutions, in separating clusters of stations with similar climatic patterns. Furthermore, in order to avoid the impact of any other external influence like altitude, anomalies were used in the analysis for both temperature and rainfall.

Based on the results of PCA (see chapter 4), different methods to calculate seasonal time-series of weighted regional averages of precipitation were explained. Two versions of weights are considered to estimate the regional series to compare the stability of the results. Standardised anomalies were also described; they smooth sudden fluctuations, while basically preserving the original climatic patterns.

Weather extreme indices are defined using the guidelines of the Expert Team (ET) on Climate Change Detection and Indices (ETCCDI). These indices are going to be calculated using the long-term and high-quality databases of rainfall and temperature

described in section 3.2. The main objective is to increase the understanding of weather extremes in Mexico; as the few studies of climate extremes were part of global or regional assessments.

Kendall's tau was selected as an alternative to the usual Pearson correlation coefficient due to its possibility of dealing with small datasets and a great number of tied ranks. This non-parametric correlation technique has better statistical properties. Because meteorological responses to large-atmospheric controls are sometimes delayed, lag cross correlations try to find the lag that maximises the coherence between two variables. In this thesis lag correlation is a method that is going to be applied to find the optimum relationships between regional precipitation averages or weather extreme indices and El Niño. This technique is preferred for its efficiency and consistent results, but is also sometimes limited to linear correlations so lacking robustness.