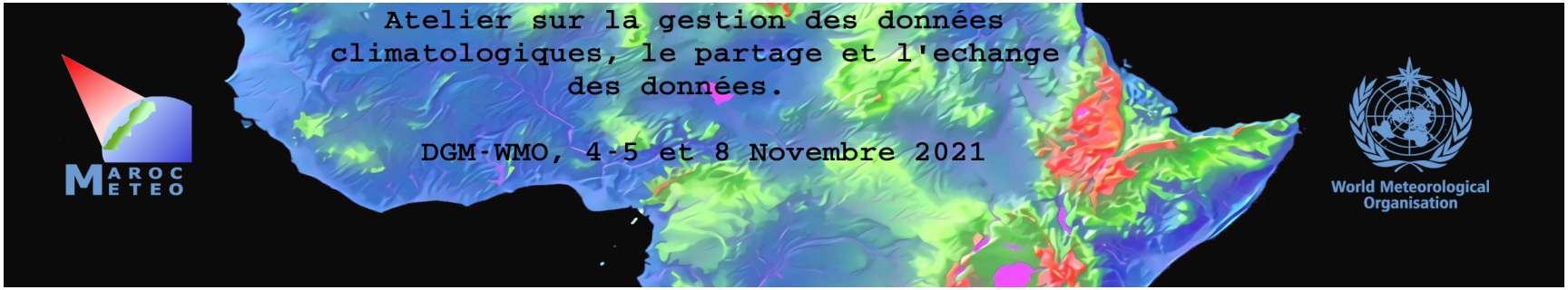


Introduction à l'évaluation de la qualité et de l'homogénéité des données

Driss BARI

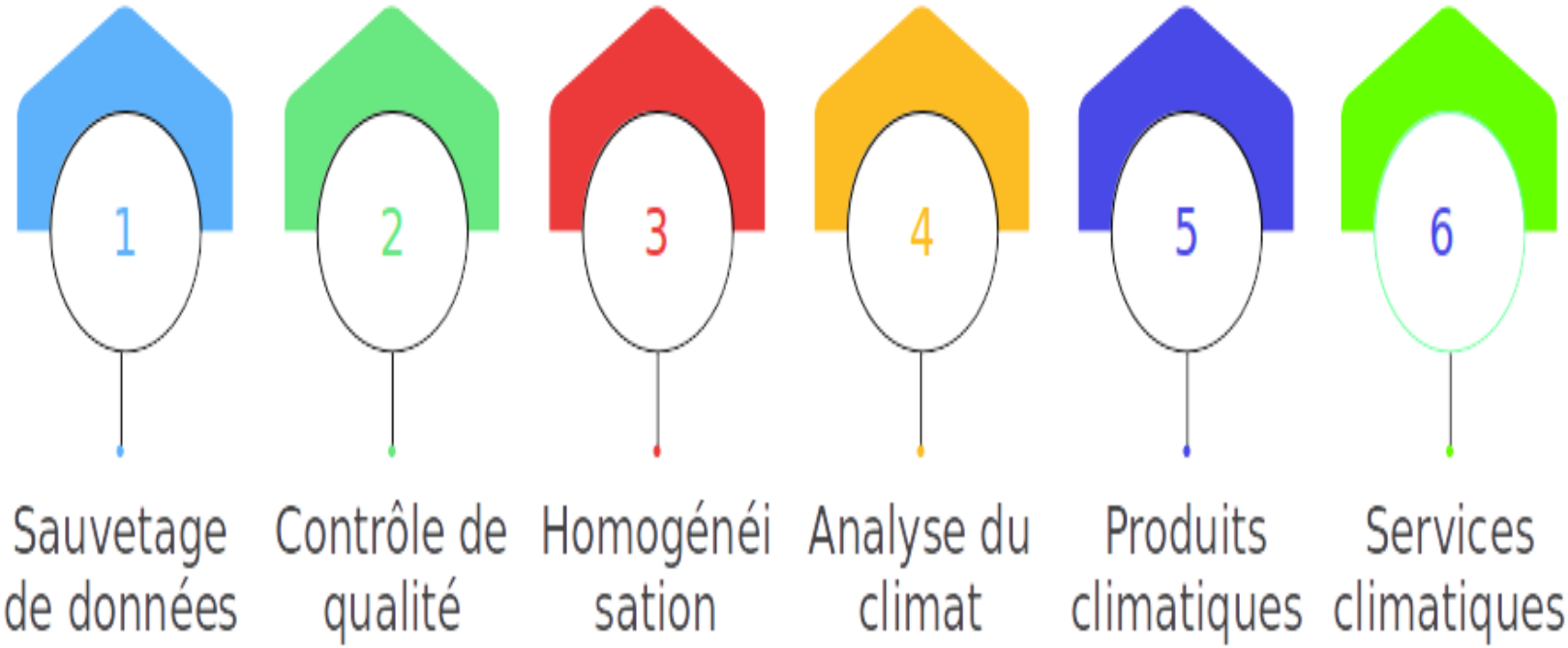
Centre National du Climat
Direction Générale de la Météorologie, Casablanca, Maroc
bari.driss@gmail.com

Atelier sur la gestion des données climatologiques, le partage et l'échange des données
DGM-OMM 4-5 et 8 Novembre 2021

Plan

Plan

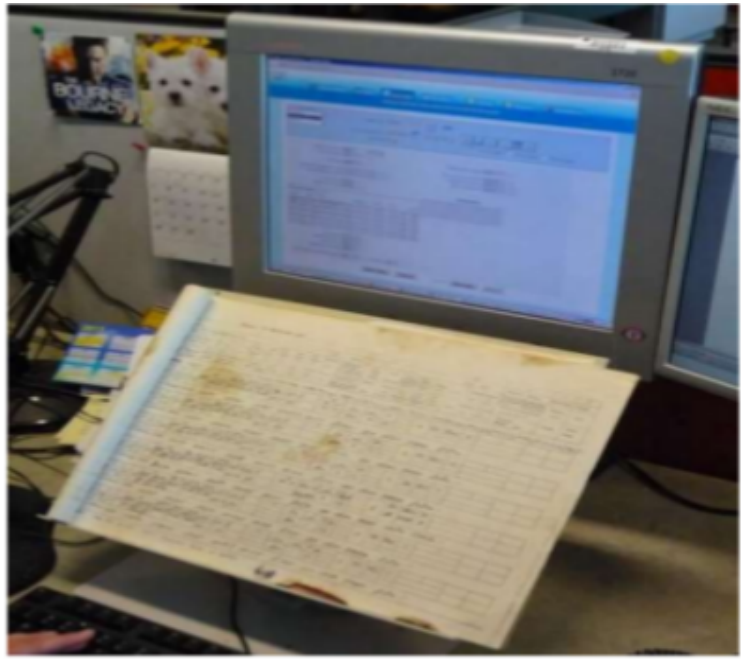
Processus

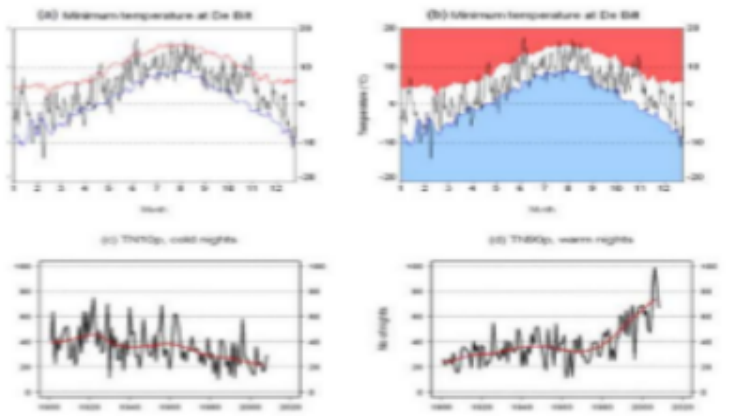

1
2
3
4
5
6
Sauvetage de données
Contrôle de qualité
Homogénéisation
Analyse du climat
Produits climatiques
Services climatiques

Processus

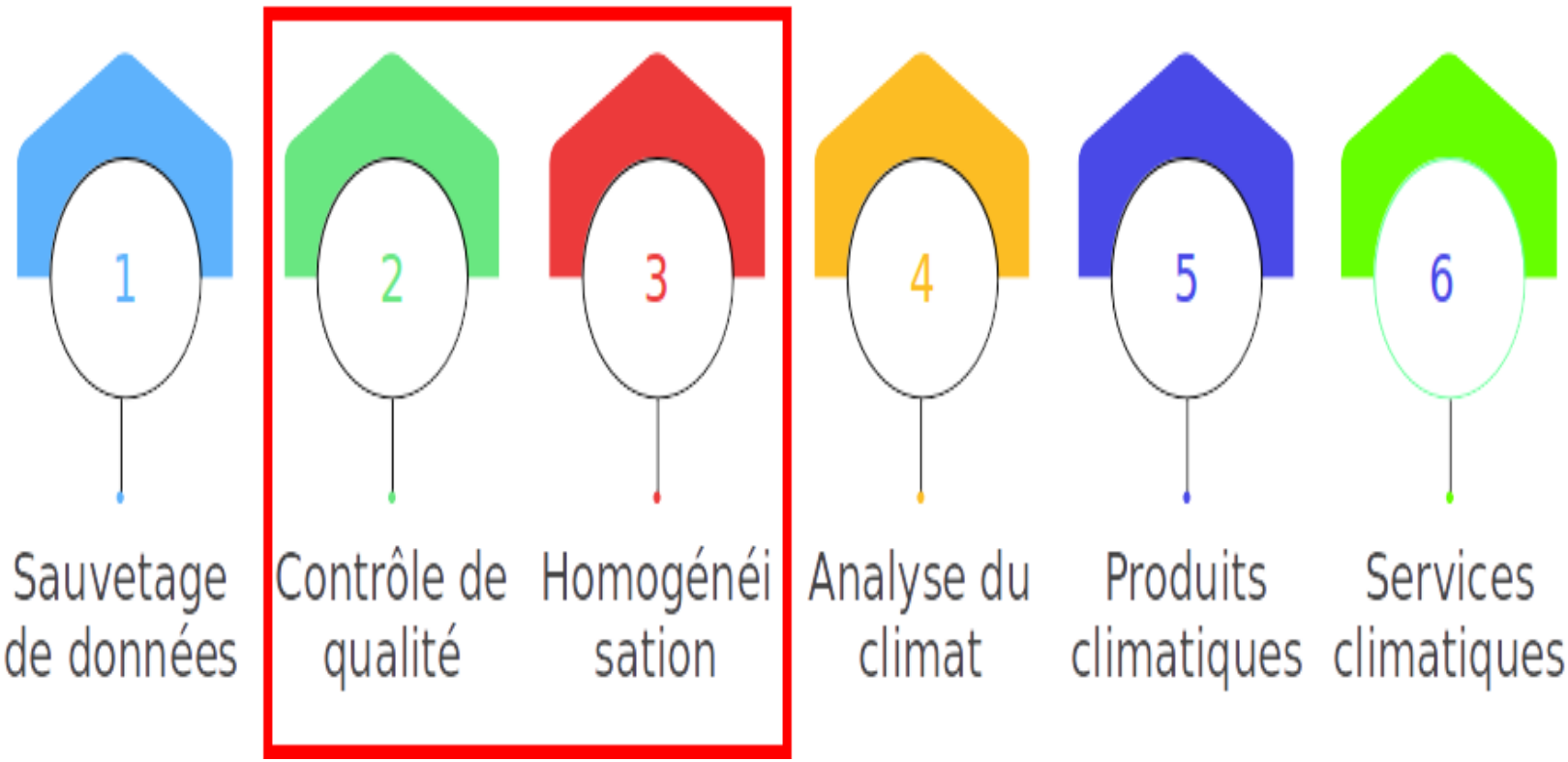
1

Sauvetage de données





Processus


1
2
3
4
5
6
Sauvetage de données
Contrôle de qualité
Homogénéisation
Analyse du climat
Produits climatiques
Services climatiques

Plan

Contrôle de la qualité des données

Les vérifications visent à déterminer **la représentativité** des données dans le temps et l'espace ainsi que leur **cohérence interne**, et à signaler les éventuelles erreurs ou incohérences.
Le contrôle qualité a pour objet de garantir que les données météorologiques et climatologiques présentent un degré de fiabilité suffisant pour les utilisateurs potentiels. Il fait donc partie du processus général d'évaluation de la qualité des données.

Contrôle de la qualité des données

Assurance de la qualité des données

Les procédures de contrôle de la qualité des données météorologiques servent en particulier à s'assurer des niveaux de qualité des données destinées aux applications et services climatologiques. **Les procédures de contrôle de la qualité appliquées devraient faire l'objet d'une documentation appropriée et être mises à la disposition des utilisateurs des données.**

Tous les détails disponibles sur les techniques exactes appliquées seront d'une grande aide pour le futur utilisateur de données, s'ils sont fournis, ainsi que des informations sur les données qui échouent aux tests et la période pendant laquelle les tests ont été réalisés.

Types d'erreurs

- **Les erreurs de données** découlent principalement d'erreurs **instrumentales**, d'erreurs **commises par l'observateur**, d'erreurs de **transmission**, d'erreurs **de saisie**, d'erreurs **de validation**, ainsi que de modifications de formes de présentation.

- **Les erreurs portant sur les métadonnées** se traduisent souvent par des erreurs de données. Si l'indicatif de la station est inexact, on pourra comprendre que la donnée provient d'un autre endroit, ou si la date est incorrecte, on pourra penser que la donnée provient d'une observation exécutée à une heure différente.

Types de tests de contrôle de qualité

- Tests des **formes de présentation** : répétitions d'observation; des dates impossibles, etc.
- Tests de **complétude** : Quand des données sont manquantes, cela peut avoir une importance cruciale suivant le type d'élément observé. Des hauteurs totales mensuelles de pluie peuvent être fortement mises en doute s'il manque quelques jours de données, en particulier si cela correspond à une période de pluie.
- Tests de **cohérence** : On distingue quatre sortes de cohérence: **interne, temporelle, spatiale et des résumés de données.**
- Test de **dispersion** : Ces vérifications établissent des **limites supérieures et inférieures** pour les valeurs possibles d'un élément climatologique (notamment la direction du vent, la nébulosité, le temps passé et le temps présent).

RClimdex et EXTRAQC

http://etccdi.pacificclimate.org/software.shtml

Expert Team (ET) on Climate Change Detection and Indices (ETCCDI)

Tests de **cohérence interne**

la cohérence interne se fonde sur des relations physiques entre les éléments climatologiques.

- s'assurer que la température du thermomètre sec est égale ou supérieure à la température du thermomètre mouillé.
- vérifier la vraisemblance de la relation entre la visibilité et le temps présent.
- la valeur maximale doit être égale ou supérieure à la valeur minimale
- La durée d'insolation par exemple se limite à la durée de la journée
- le rayonnement global ne peut être plus grand que l'éclairement énergétique au sommet de l'atmosphère
- la direction du vent doit se situer entre 0°et 360°
- les hauteurs de précipitations ne peuvent être négatives

Tests de **cohérence temporelle**

la cohérence temporelle elle vérifie la variation d'un élément dans le temps. La variation est généralement fonction de l'élément, de la saison, du lieu et de l'intervalle de temps écoulé entre deux observations successives.

- mettre en doute une baisse de la température de 10 °C en une heure (bien que cela soit fort réaliste dans le cas du passage d'un front froid ou de l'apparition d'une brise de mer)
- Pour certains éléments, une absence de variation peut indiquer une erreur. Une série de vitesses du vent identiques peut par exemple indiquer un problème d'anémomètre.

Tests de **cohérence spatiale et des résumés de données**

la cohérence spatiale

Pour vérifier la cohérence spatiale, on compare chaque observation aux observations exécutées à la même heure à d'autres stations de la région.

la cohérence des résumés de données

En comparant différents résumés de données, il est possible de déceler les erreurs portant sur des valeurs individuelles ou sur un résumé.

Par exemple, la somme et les moyennes des valeurs quotidiennes peuvent être calculées pour différentes périodes (une semaine, un mois, une année).

Dans le cas d'un élément comme la hauteur de pluie dont la mesure représente un cumul, il suffit d'effectuer un recoupement entre la somme des douze mois et la somme de toutes les valeurs quotidiennes enregistrées au cours de l'année correspondante pour déceler une erreur.

EXTRAQC

Les routines **EXTRAQC** sont un ensemble de fonctions codées R pour le contrôle qualité. Elles sont développés par Enric Aguilar et Marc Prohom (Espagne) et ont été intégrées dans le logiciel RClimdex développé par l'ETCCDI.

Les routines EXTRAQC incluent les tests suivants:

- Contrôle des dates en double
- Évaluation des problèmes d'arrondi
- Valeurs hors limites, basées sur des valeurs de seuil fixes
- Valeurs aberrantes, basées sur le dépassement de la plage interquartile
- Différences interdiurnes basées sur des valeurs seuils fixes
- Cohérence entre les températures maximales et minimales (Tmax > Tmin)
- Contrôle des valeurs égales et consécutives

EXTRAQC

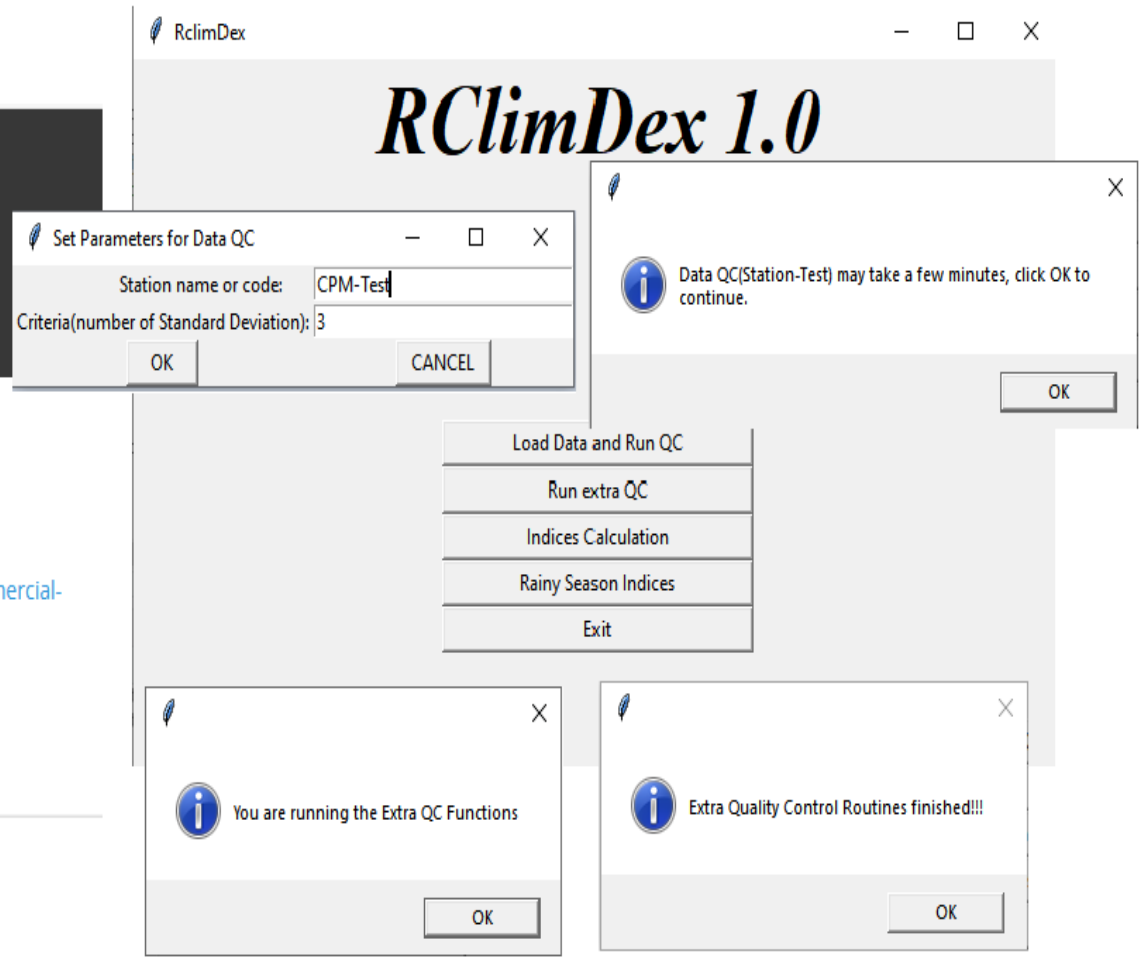
http://www.c3.urv.cat/softdata.php

Q.C Software

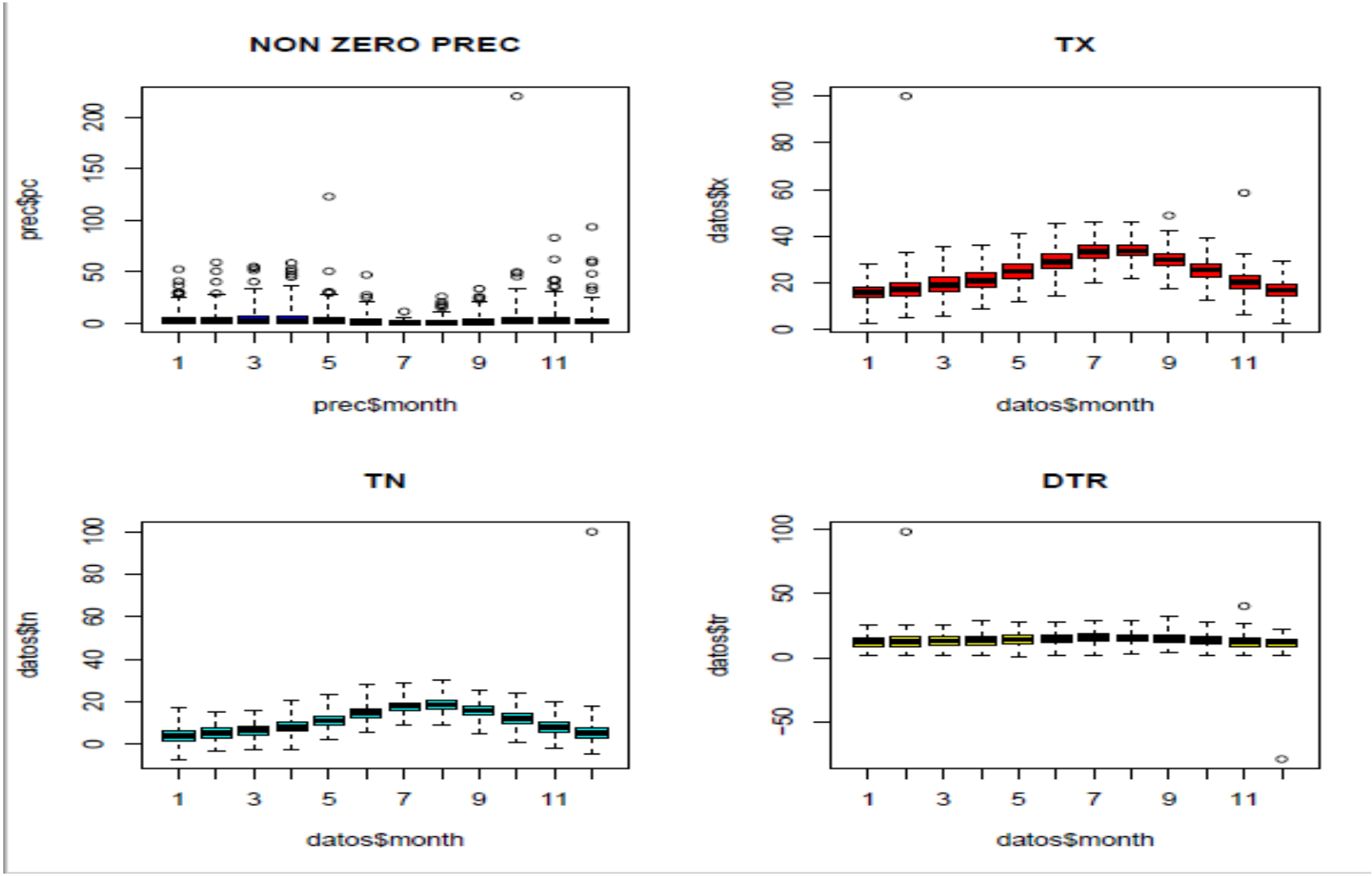
RClimDex-extraqc

Manual

Download



EXTRAQC : exemple



EXTRAQC : exemple

Ocurrence of 4 or more equal consecutive values

2001	5	15	0	26.7	8.7
2001	5	16	0	28.7	12.4
2001	5	17	0	29.8	14.3
2001	5	18	0	26.1	13.5
2001	5	19	0.9	15.5	13.6
2001	5	20	0	22	8.6
2001	5	21	0	27.2	8.8
2001	5	22	0	27.2	10.4
2001	5	23	0	27.2	11.5
2001	5	24	0	27.2	9.5
2001	5	25	0	27.2	10.9
2001	5	26	0	27.2	11.8
2001	5	27	0	27.2	13.7
2001	5	28	0	33.2	12.6
2001	5	29	0	28	13
2001	5	30	0	30.5	13.9

année mois jour RR TX TN

-99.9 indique valeur manquante

Jumps : the temperature difference with the previous day is greater or equal than 20 °C

2015	12	2	0	19.4	4.1
2015	12	3	0	20.6	99.9
2015	12	4	0	19.4	5.9
2015	12	5	0	19.6	3.1
2015	12	6	0	19.7	3.7
2015	12	7	0	19.5	5.8

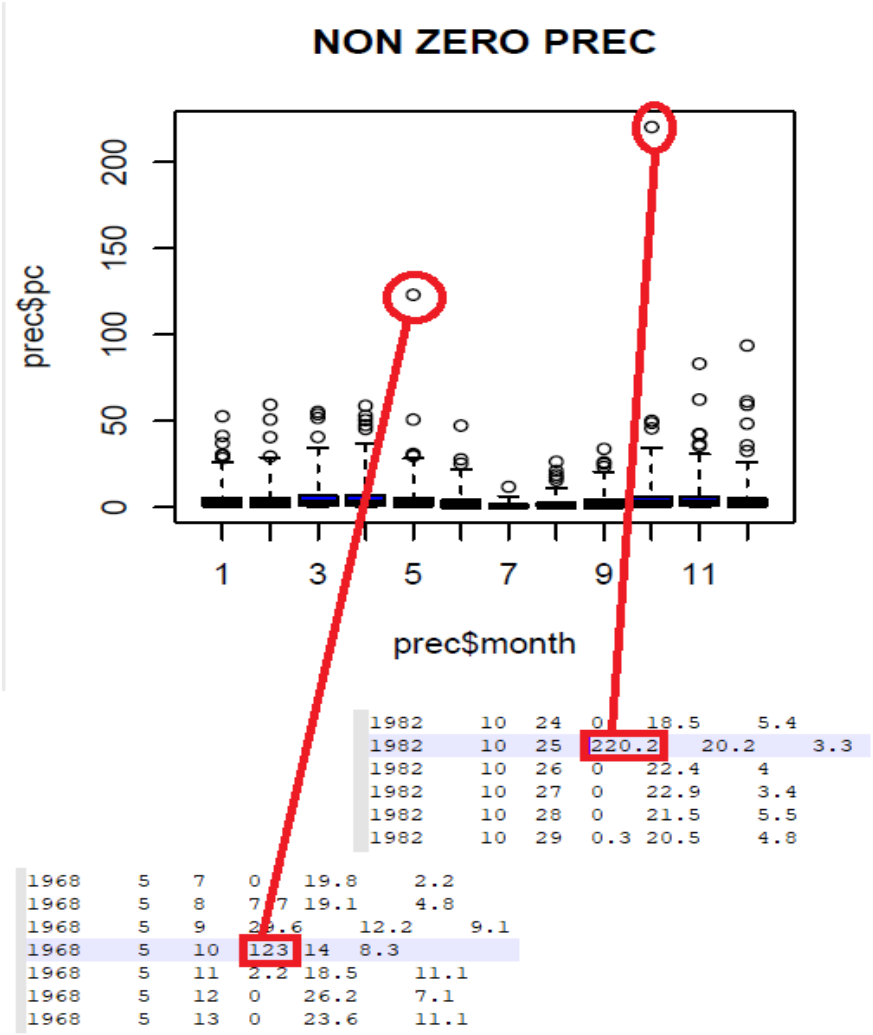
Maximum temperature is lower than minimum temperature

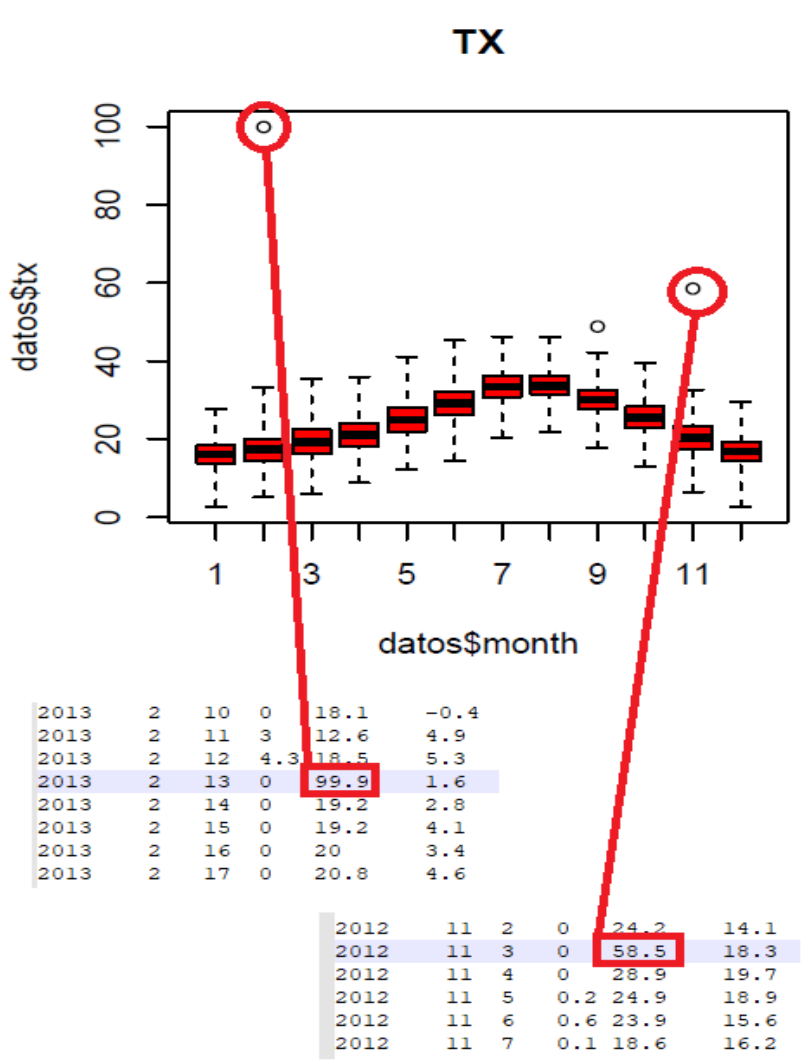
2015	12	2	0	19.4	4.1
2015	12	3	0	20.6	99.9
2015	12	4	0	19.4	5.9
2015	12	5	0	19.6	3.1

Too large : precipitation values exceeding 200 mm and temperature values exceeding 50 °C.

1982	10	25	220.2	20.2	3.3
2012	11	3	0	58.5	18.3
2013	2	13	0	99.9	1.6
2015	12	3	0	20.6	99.9

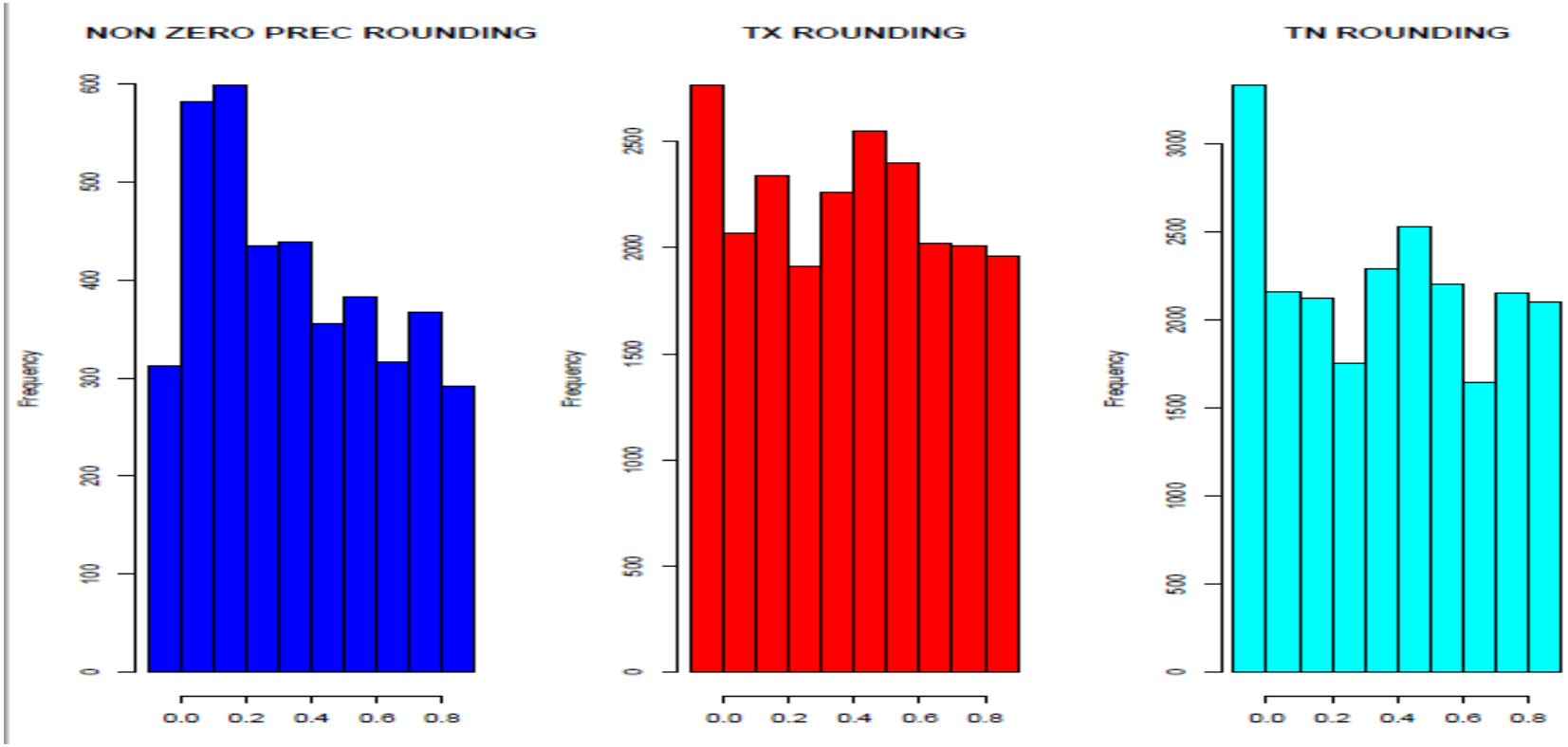
EXTRAQC : exemple





EXTRAQC : exemple

Il examine les problèmes d'arrondi en traçant les valeurs après la virgule décimale. Il montre à quelle fréquence chacune des 10 valeurs possibles (.0 à .9) apparaît. On s'attend à ce que toutes ces valeurs soient représenté.



Plan

homogénéité des données

Les données climatiques peuvent fournir de nombreuses informations sur l'environnement atmosphérique qui ont un impact sur presque tous les aspects de l'activité humaine. L'analyse du climat repose sur des séries chronologiques longues. Si l'on veut évaluer si tel ou tel endroit s'est réchauffé ou est devenu plus humide, il faut examiner 50, 60, ... 100 ans de données. **Cependant**, pour que ces analyses et d'autres analyses climatiques à long terme soient précises, les données climatiques utilisées doivent être aussi homogènes que possible.

Une série chronologique climatique homogène est définie comme une série où les variations ne sont causées que par des variations climatiques.

Les principales causes des discontinuités

Malheureusement, la plupart des séries chronologiques climatologiques à long terme **ont été affectées par un certain nombre de facteurs non climatiques** qui rendent ces données non représentatives de la variation climatique réelle se produisant au fil du temps.

Ces facteurs comprennent des changements dans :

- les instruments et/ou abri météo,
- les pratiques d'observation (Changement d'observateur et/ou heures d'observation),
- les emplacements des stations,
- les formules utilisées pour calculer certains paramètres,
- l'environnement de la station.

Processus d'homogénéisation des données

Certains changements **induisent de fortes discontinuités** tandis que d'autres changements, en particulier des changements dans l'environnement autour de la station, peuvent **induire des biais graduels** dans les données. Toutes ces inhomogénéités peuvent biaiser une série chronologique et **conduire à des interprétations erronées du climat étudié**. Il est donc important **de supprimer les inhomogénéités ou au moins de déterminer l'erreur possible qu'elles peuvent provoquer.**

Homogénéisation

Technique consistant à rendre les séries chronologiques homogènes, en appliquant des méthodes statistiques scientifiquement solides pour éliminer les effets de biais artificiels, tels que ceux causés par des changements dans les pratiques d'observation, l'instrumentation, l'emplacement, etc.

Contexte de l'homogénéité

L'homogénéité temporelle d'un ensemble de données climatiques est essentielle dans la recherche climatologique, en particulier lorsque les données sont utilisées pour valider des modèles climatiques ou pour évaluer le changement climatique et ses impacts environnementaux et socio-économiques associés. Par conséquent, **il serait essentiel de signaler si des tests d'homogénéité ont été appliqués aux données.**

- Quels éléments ont été testés pour l'homogénéité ?
- Pendant quelles périodes ?
- Sur quelle échelle de temps (quotidienne, mensuelle, saisonnière ou annuelle) ?
- Nombre d'inhomogénéités trouvées dans la série chronologique.
- etc.

Contexte de l'homogénéité

Lors de l'évaluation de l'homogénéité de la série, nous essayons d'identifier à l'aide de techniques statistiques et de métadonnées où l'hétérogénéité de la série a été rompue et nous essayons d'ajuster l'effet de ces ruptures, pour améliorer la qualité de notre inférence climatique.

Évaluation de l'homogénéité des données climatiques

Il est presque impossible d'être sûr à 100% de la qualité du passé données, une évaluation de l'homogénéité est toujours recommandée. Il n'y a pas une seule meilleure technique à recommander. Cependant, les quatre étapes énumérées ci-dessous sont généralement suivies :

1. Analyse des métadonnées et contrôle qualité

même en présence des métadonnées les plus soigneusement documentées, il est conseillé de comparer ce que dit l'historique de la station et ce que l'analyse des données identifie, comme une sorte de double contrôle.

Évaluation de l'homogénéité des données climatiques

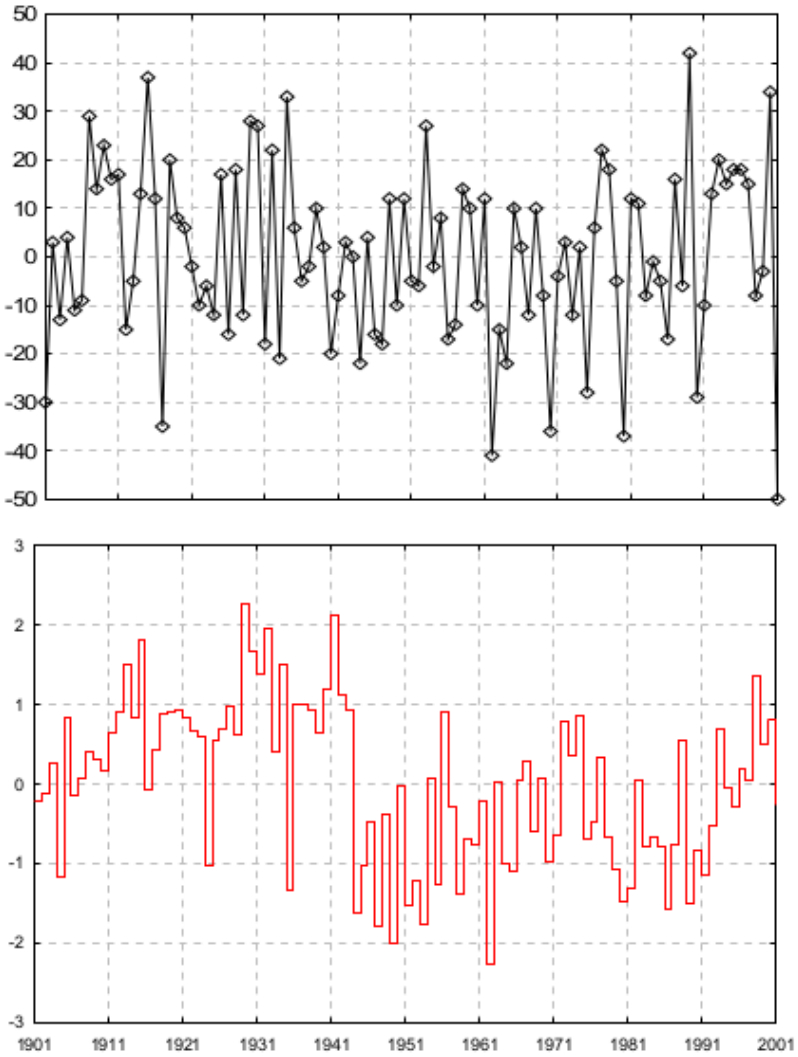
2. Création d'une série chronologique de référence

- utilise une homogénéisation relative, c'est-à-dire que nous comparons des séries chronologiques avec celles des stations avoisinantes bien corrélés entre elles.
- Cette comparaison peut être effectuée sous forme de comparaisons par paires, de moyennes de plusieurs stations ou de méthodes plus sophistiquées, telles que les composantes principales (ACP).

3. Détection des ruptures

- L'idée est de créer des différences (ex. température) ou des rapports (ex. précipitation) d'une série candidate (celle que l'on veut homogénéiser) vers une référence
- La série candidate et la référence partagent le même climat donc les caractéristiques étranges de la différence entre les deux séries ne sont pas dues à l'évolution du climat.

Détection des ruptures



En haut : Moyenne mensuelle des températures minimales quotidiennes pour décembre à Burgos, Espagne. Données en 1/10 °C ;

En bas : différence entre les séries chronologiques candidate et de référence normalisées calculées à la suite du test d'homogénéité normale standard, en utilisant 10 stations voisines.

La différence entre les séries chronologiques candidate et de référence (en bas) montre clairement une inhomogénéité en 1941, documentée dans les métadonnées comme une relocalisation.

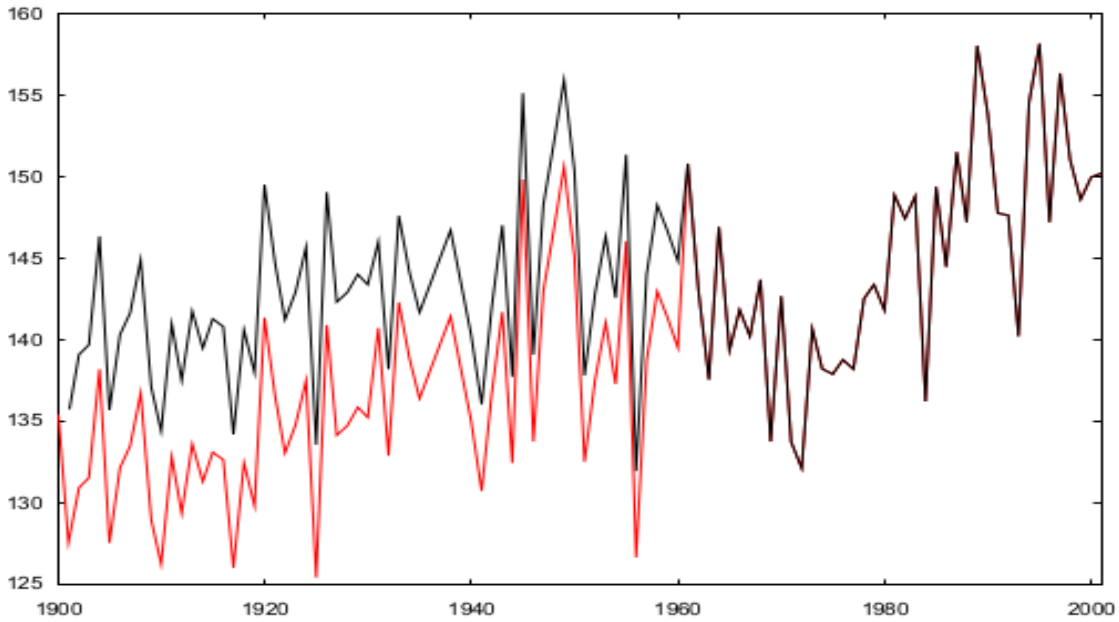
Les données d'origine (en haut) masquent l'inhomogénéité.

Source : GUIDELINES ON CLIMATE METADATA AND HOMOGENIZATION. Enric Aguilar et al. 2003. WMO-TD No. 1186

4. Ajustement des données

- Une fois l'identification des ruptures est terminée, l'étape suivante consiste à décider quelles ruptures seront acceptés comme de réelles inhomogénéités.
- L'ajustement des données est la correction appliquée aux données pour améliorer leur homogénéité et rendre toutes les observations comparables aux dernières données disponibles.
- Il est toujours recommandé de corriger les données pour qu'elles correspondent aux conditions de sa section homogène la plus récente.

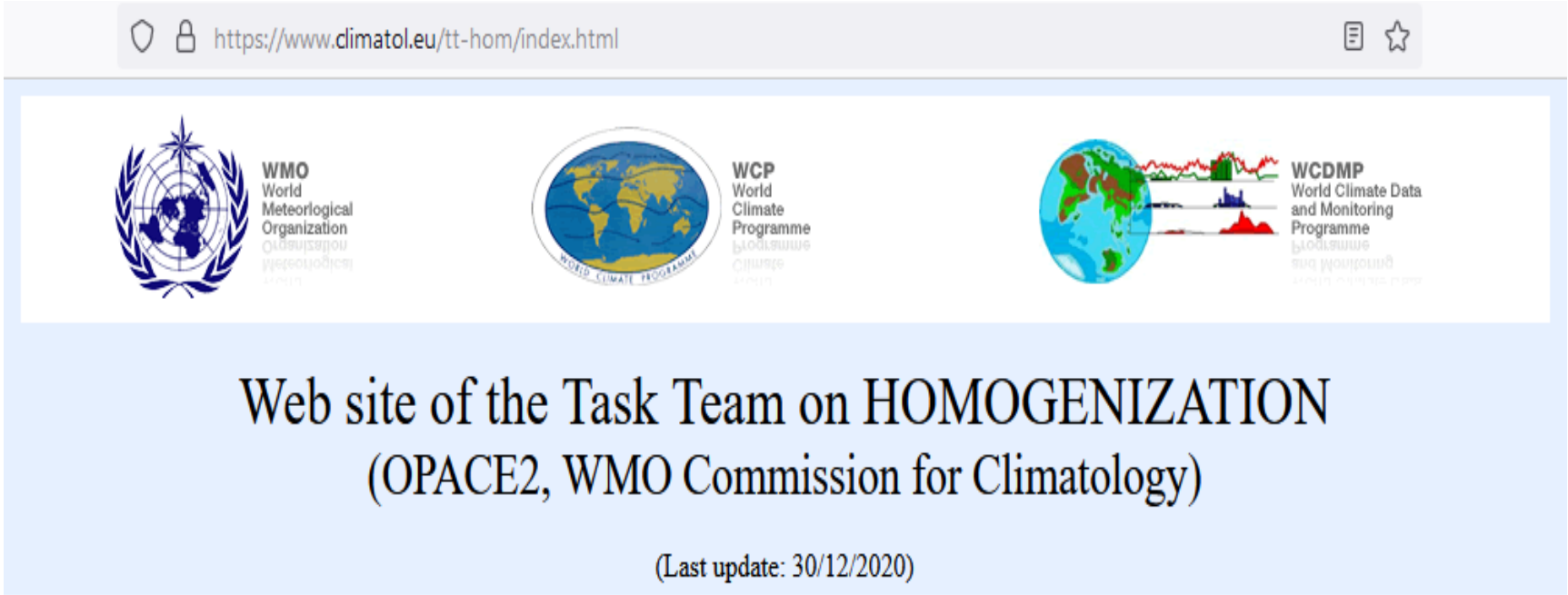
Homogeneity assessment for climate data



Moyennes annuelles originales (ligne rouge) et ajustées (ligne noire) de la température moyenne quotidienne pour Madrid, Espagne. Données au 1/10°C. Les données ont été ajustées pour les changements soudains tendances moyennes et artificielles à l'aide d'un test itératif qui compare la valeur moyenne de deux périodes différentes sur une série chronologique de référence standardisée, calculée à partir d'un certain nombre de stations de référence bien corrélées. Les données inhomogènes (ligne rouge) montrent une tendance beaucoup plus large pour la période de 100 ans, car elles contiennent de véritables fluctuations climatiques ainsi que des biais artificiels.

Source : Aguilar, E (2002) « Homogenizing the Spanish Temperature Series ».

Logiciels d'homogénéisation



https://www.climatol.eu/tt-hom/index.html

Logiciels d'homogénéisation

Présentation générale des caractéristiques des logiciels d'homogénéisation

Logiciel	*Résolution (détection)*[a]	*Méthode de détection*	*Utilisation de références*[b]	*Résolution (correction)*[c]	*Méthode de correction*	*Opération principale*	*Utilisation de métadonnées*	*Variable*[d]	*Documentation*[e]	*Référence*
ACMANT[f]	Année, mois	Points de rupture multiples	Composite	Année, mois, [jour]	Conjointe (ANOVA)	Automatique	Non	Toute	Guide d'utilisation	Domonkos et Coll (2017)
AnClim	Année, mois	Plusieurs	Composite, par paires	Année, mois, jour	Plusieurs	Interactive, automatique	Oui	Toute	Manuels	Štěpánek *et al.* (2009)
Berkeley Earth	Mois	Division	Composite	s/o	s/o	Automatique	Oui	T	Article	Rohde *et al.* (2013)
Climatol	Mois (série), jour	Division	Composite	Mois (série), jour	Comblement des données manquantes	Automatique	Oui	Toute	Manuel et guide d'utilisation	Guijarro (2018)
GAHMDI HOMAD	Mois (série), jour	Points de rupture multiples	Sélection	Jour	Méthode des moments d'ordre supérieur	Automatique	Oui	T	Aucune	Toreti *et al.* (2010, 2012)
GSIMCLI	Année, mois	Points de rupture multiples	Composite	Voir note de bas de page [g]	Voir note de bas de page [g]	Automatique et interactive	Non	T, p	Manuels	Ribeiro *et al.* (2017), Costa et Soares (2009)
HOMER	Année, saison, mois	Points de rupture multiples	Par paires, conjointe	Année, mois	Conjointe (ANOVA)	Interactive	Oui	Toute	Guide d'utilisation de base + cours	Mestre *et al.* (2013)
iCraddock	Année, saison, mois	Division	Par paires	Année, saison, mois, jour	Quotidien-nement: corrections mensuelles lissées	Interactive	Oui	Toute	Aucune	Craddock (1979), Brunetti *et al.*, 2006
MASH	Année, saison, mois	Points de rupture multiples	Composite	Mois, [jour]	Comparaisons multiples	Automatique et interactive	Oui	Toute	Guide d'utilisation	Szentimrey (2008, 2014)
ReDistribu-tion Test	Relevés	Points de rupture uniques	Pas de référence	s/o	s/o	Interactive	Non (mais interactif)	Vent	Aucun	Petrovic (2004)
RHtests	Année, mois, jour	Division	Sélection ou pas de référence	Année, mois, jour	Régression multiphase	Interactive	Oui	Toute	Guide d'utilisation + cours	Wang (2008a et b), Wang et Feng (2013)

Logiciels d'homogénéisation

Package	Version	License	Open source	Operating System	Program type	Primary operation	Availability
ACMANT	4	Freeware	No	DOS/Windows	Executable	Automatic	https://github.com/dpeterfree/ACMANT
AnClim ProClimDB	?	Freeware	No	Windows	Executable	Interactive (and automatic)	https://www.climahom.eu/
Climatol	3.0	GPL	Yes	(Most)	R package	Automatic	https://www.climatol.eu/index.html
GAHMDI HOMAD	?	GPL	Yes	(Most)	R source R/Fortran	Automatic Interactive	mail to andrea.toreti at giub.unibe.ch
GSIMCLI	0.0.1	GPL	Yes	(Most)	Python	Automatic (and interactive)	https://iled.github.io/gsimcli/
HOMER	2.6	GPL	Yes	(Most)	R source	Interactive	https://www.climatol.eu/pub/HOMER2.6.zip
MASH	3.03	Freeware	No	DOS/Windows	Executable	Automatic (and interactive)	https://www.met.hu/en/omsz/rendezvenyek/homogenization_and_interpolation/software/
ReDistribution Test	?	Freeware	Yes	(Most)	R source	Interactive	mail to predrag.petrovic at hidmet.gov.rs
RHtests	4	Freeware	Yes	(Most)	R source	Interactive	https://etccdi.pacificclimate.org/software.shtml
USHCN	52i	Freeware	Yes	Some linux versions	Fortran source	Automatic	ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/v3/software/52i/phav52i.tar.gz

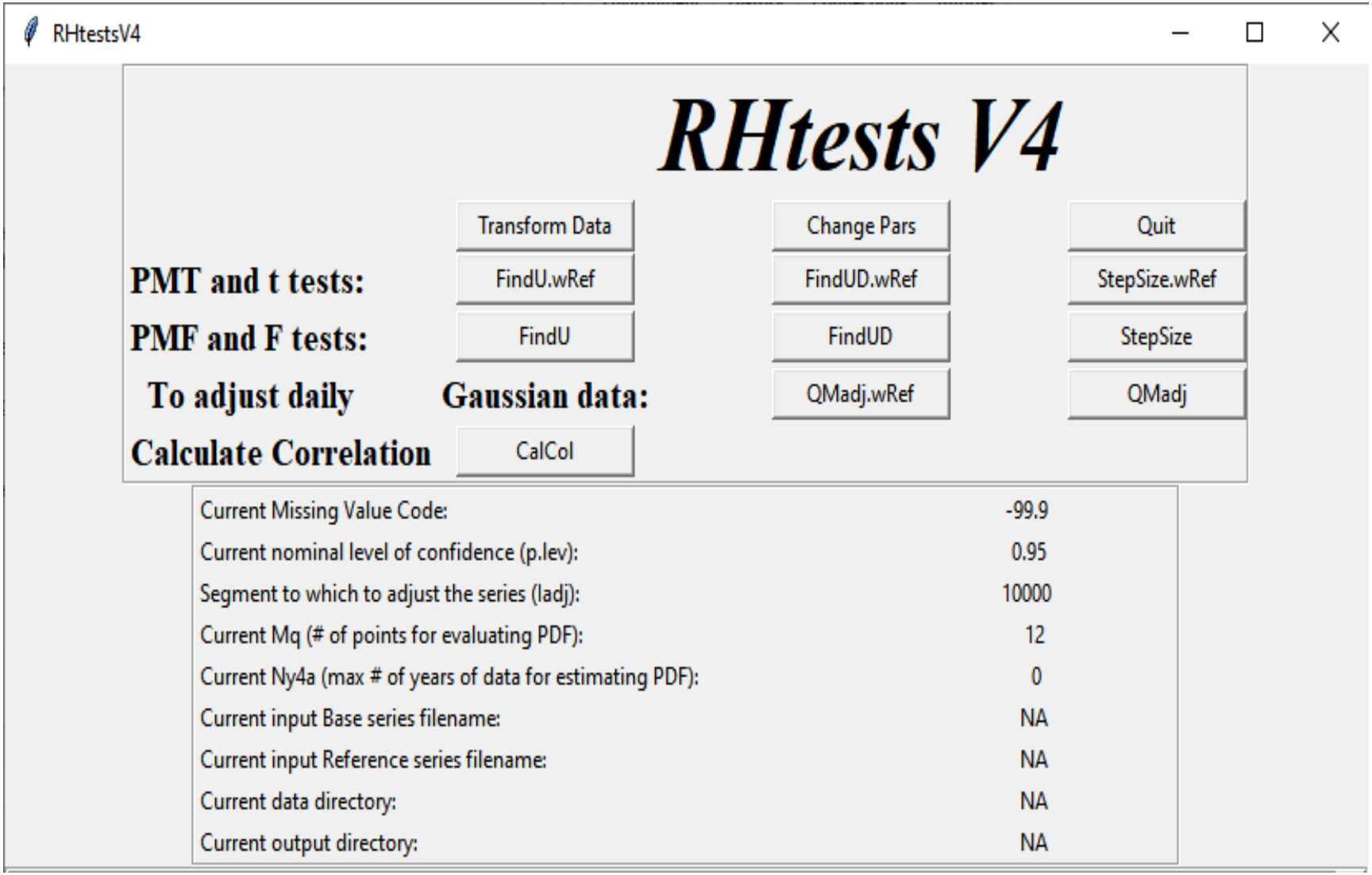
Logiciels d'homogénéisation

Package	GUI	Time resolution	Input format	Metadata use	Detection method	Ref. series selection	Detection statistic	Climatic variables
ACMANT	No	Monthly & daily	ASCII	No	Reference	Correlation	Caussinus-Lyazrhi	Temperature and precipitation
AnClim ProClimDB	Yes	Any	ASCII DBF	Yes	Ref. and pairwise	Correlation & distance	Several	Any
Climatol	No	Monthly & daily	ASCII	Yes	Reference	Distance	SNHT	Any
GAHMDI HOMAD	No	Monthly Daily	ASCII	Yes	Pairwise	Correlation	New method	Any Temperature
GSIMCLI	Yes	Monthly & yearly	ASCII	No	Multiple references	Correlation & distance	User defined	Any
HOMER	No	Monthly	ASCII	Yes	Pairwise	Correlation	Penalized Likelihood	Any
MASH	No	Monthly & daily	ASCII	Yes	Multiple references	Correlation	MLR & Hypothesis test	Any
ReDistribution Test	No	Sub-daily	ASCII	No	Distribution	None	SNHT-like	Wind speed and direction
RHtests	Yes	Monthly & daily	ASCII	Yes	Reference	Correlation	Penalized max. t & F tests	Any
USHCN	No	Monthly	ASCII	Yes	Pairwise	Correlation	MLR	Temperature

Logiciels d'homogénéisation

				Outputs				
Package	Correction method	Missing data tolerance	Max. number of series	Homogenized series	Corrected outliers	Corrected breaks	Graphics	Documentation
ACMANT	ANOVA	Very high	4000	Yes	Yes	Yes	No	User's guide
AnClim ProClimDB	Several	User defined	?	Yes	Yes	Yes	Yes	Manuals
Climatol	Missing data filling	Very high	9999*	Yes	Yes	Yes	Yes	User's guide
GAHMDI HOMAD	?	?	?	Yes	No	Yes	Yes	None
GSIMCLI	User-defined & missing data filling	High	9999*	Yes	Yes	Yes	No	Manuals
HOMER	ANOVA	15 year data	?	Yes	Yes	Yes	Yes	User's guide
MASH	Multiple comparisons	30%	500	Yes	Yes	Yes	Yes	User's guide
ReDistribution Test	None	10-20%	?	No	No	Detected breaks	No	None
RHtests	Multi-phase regression	?	1	Yes	No	Yes	Yes	User's guide
USHCN	Multiple comparisons	Very high	9999*	Yes	?	Yes	No	Plain text notes

Logiciel RHtests V4



Plan

Documents de référence

Organisation Météorologique Mondiale

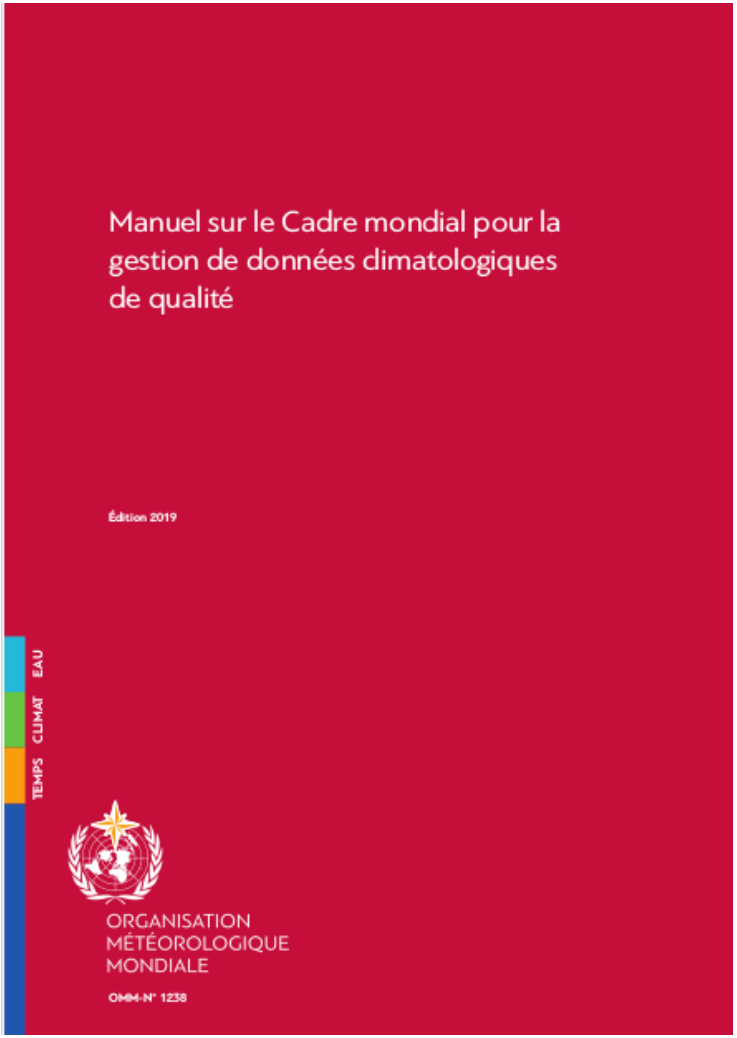
WMO No. 100

Guide des pratiques climatologiques

Edition 2018

`https://library.wmo.int/doc_num.php?explnum_id=9864`

Documents de référence



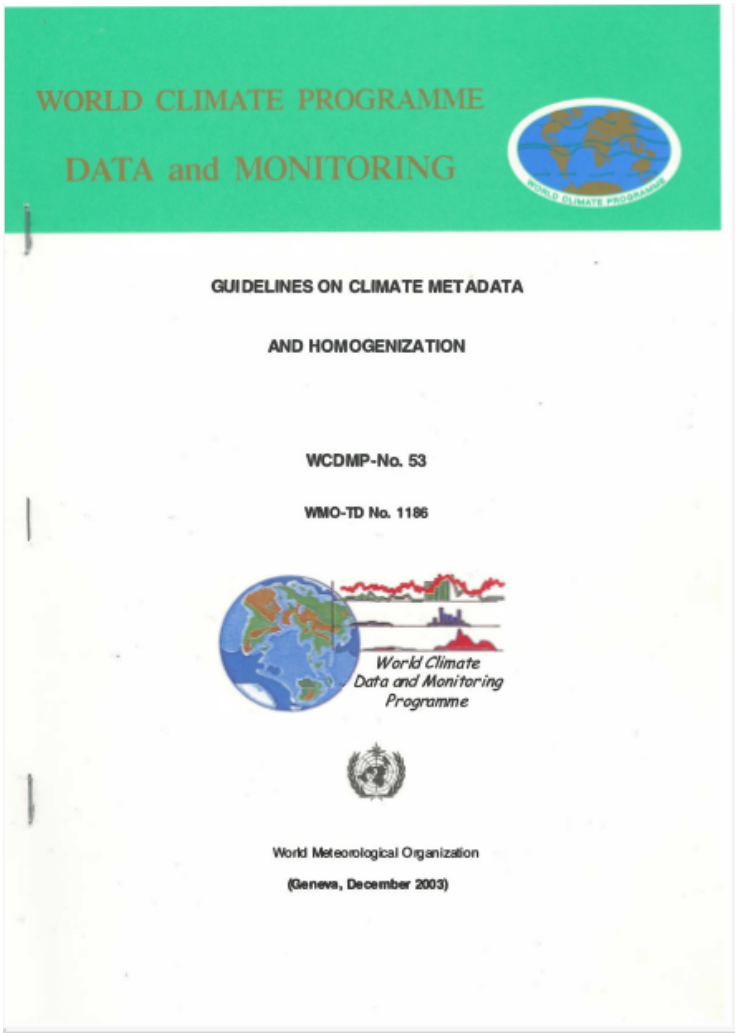
Organisation Météorologique Mondiale

WMO No. 1238

Manuel sur le Cadre mondial pour la gestion de données climatologiques de qualité

Edition 2019

Documents de référence



Organisation Météorologique Mondiale

WMO-TD No. 1186

Directives sur les métadonnées climatique et l'homogénéisation (En Anglais seulement)

Edition 2003

Documents de référence

Organisation Météorologique Mondiale

WMO No. 1245

Directives sur l'homogénéisation

Edition 2020

`https://library.wmo.int/doc_num.php?explnum_id=10374`

MERCI

Driss BARI
bari.driss@gmail.com