ReKliEs-De

Nutzerhandbuch

ReKliEs-De
Regionalisiertes Klimamodellensemble für Mitteleuropa

Langfassung des Kapitels 8 „Datenverarbeitung"

Finalisierung diese Dokuments: Dezember 2017

Satz und Layout

Arne Spekat
Climate & Environment Consulting Potsdam GmbH
Dennis-Gabor-Strasse 2, 14469 Potsdam

8 Datenverarbeitung

Vorbemerkung: Dies ist die Langfassung des Kapitels 8 des Benutzerhandbuchs. In der gedruckten Fassung sowie deren pdf-Abbild ist dieses Kapitel verkürzt enthalten. Viele technische Details sind in diesem Dokument zu finden, das im Supplement enthalten ist.

Für eine Vielzahl von Anwendungen ist der Umgang mit Daten aus den Klimamodellen nicht notwendig, da es Webportale und andere Software zur Analyse und Visualisierung von Klimamodell-Ergebnissen gibt, z.B. IDP[1)] und climate4impact (https://climate4impact.eu/).

Wer selbst an die Akquise von Modelldaten herangehen möchte, sollte darüber verfügen: (i) Geduld, (ii) Findigkeit und (iii) viel freien Speicherplatz. Weder der Daten-Download, noch die Weiterverarbeitung ist selbsterklärend und die zur Verfügung stehenden Datenmengen sind immens. Bei hoch auflösenden Regionalmodellen können es für Mitteleuropa 10,000 Gitterpunkte sein, an denen zahlreiche meteorologische Variable in einer zeitlichen Auflösung z.B. von Tagen archiviert sind. Damit eine solche Datenmenge überhaupt speicher- und transferier bar wird, erfolgte die Einigung auf ein binäres, hochkomprimiertes Datenformat: netCDF (Network Common Data Form). Die Bearbeitung und die Analyse von netCDF-Daten werden etwas einfacher, wenn auf Nutzerseite das Betriebssystem UNIX oder Linux verwendet wird. Windows-Nutzer erwarten Einschränkungen.

[1)] Bitte beachten: Die Abkürzung IDP wird an dieser Stelle für das bei vielen Nutzern eingesetzte und von CEC-Potsdam GmbH erstellte Programm benutzt; im Kontext mit Open Source Network-Datenzugängen kann IDP auch Identity Provider bedeuten.

8.1 Akquise von Modelldaten

8.1.1 Datenquelle ESGF

8.1.1.1 Datensuche und Datendownload

Als eine der wichtigsten Datenquellen für Klimamodell- und Klimabeobachtungsdaten kann das ESGF (Earth System Grid Federation, siehe http://esgf.llnl.gov) angesehen werden. Im ESGF sind Daten der wichtigsten nationalen wie internationalen Klimaprojekte zu finden. Der Download ist kostenfrei und die Nutzung für wissenschaftliche Zwecke erlaubt.

Im Folgenden werden die notwendigen Schritte beschrieben, wie ein Nutzer über ein ESGF Portal Daten suchen und herunterladen kann.

a) Registrierung an einem beliebigen Portal, z.B. beim DKRZ (https://esgf-data.dkrz.de)

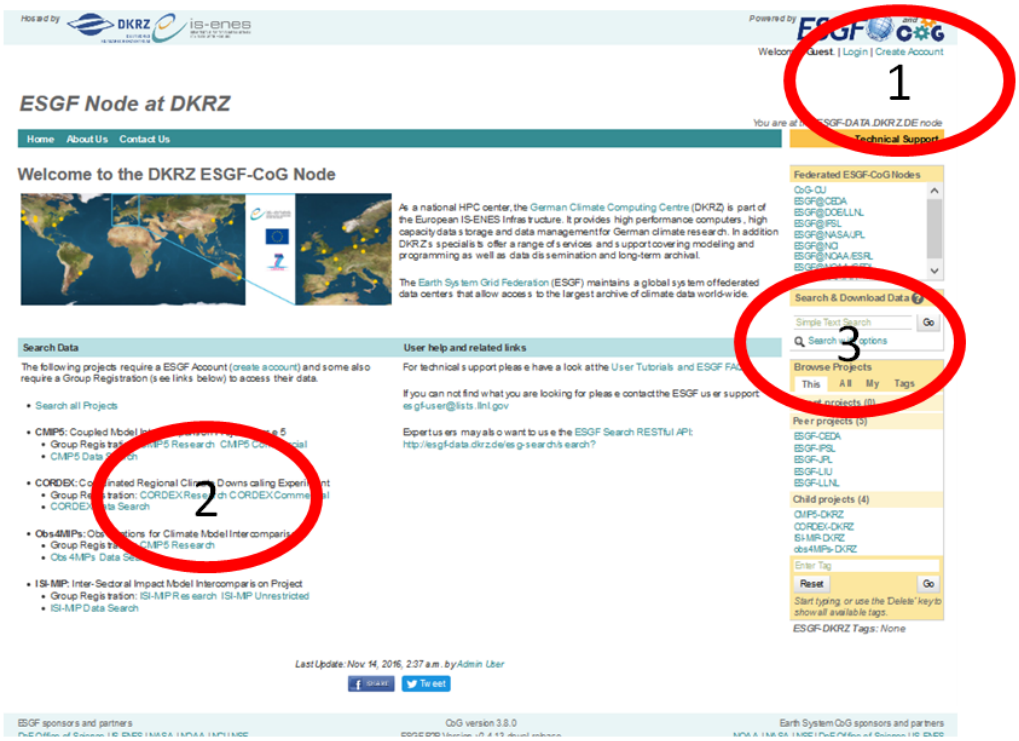

Figure 8.1: Willkommen-Bildschirmmaske des ESGF-Knotens am DKRZ. Die rot markierten Zahlen beziehen sich auf die im Text beschriebenen Schritte 1 bis 3.

Weitere Einschränkung der Suche über das Drop-Down-Menü im linken Bereich z.B.:

Schritt 1: Nutzer-Registrierung („Create Account"):
Nutzername, Password und einige Angaben zur Person
Der Nutzer bekommt einen Identifikations-Namen (OpenID):
Syntax: https://esgf-data.dkrz.de/esgf-idp/openid /Nutzername

Schritt 2: Registrierung bei einer Datenzugriffs-Kontrollgruppe („Group Registration"):
Zur Erlangung der Berechtigung, auf Daten eines Projektes zugreifen zu dürfen (Autorisierung)
Für ReKliEs-Daten wird die Mitgliedschaft in einer der beiden CORDEX-Datenzugriffs-Kontrollgruppen benötigt, für wissenschaftliche Nutzung ist am besten CORDEX Research.
Alternative Autorisierungsmöglichkeiten:

- Zugriffsversuch mit dem Browser auf ein einzelnes File dieses Projektes. Falls noch keine Autorisierung für das Projekt existiert, wird man automatisch auf eine Registrierungsseite weitergeleitet.
- Über die Liste der Datenzugriffs-Kontrollgruppen (https://esgf-node.jpl.nasa.gov/ac/list/)

b) Datensuche an einem beliebigen Portal z.B. beim DKRZ
Schritt 3: Die Suche von Daten erfolgt über das Feld ‚Search & Download Data'. Suchoptionen sind unter dem Link ‚More search options' zu finden

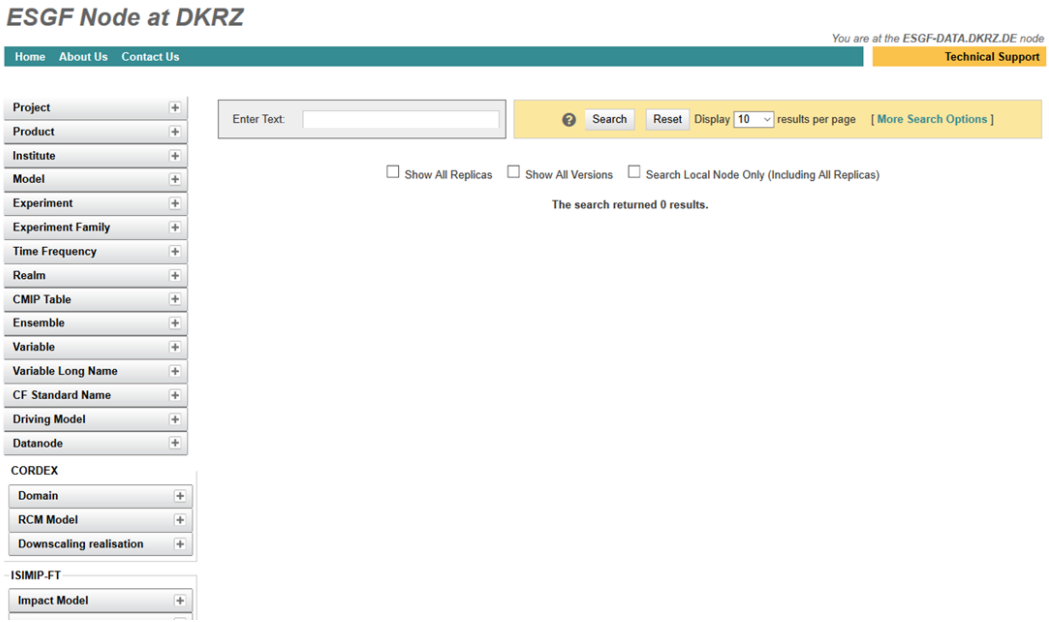

Figure 8.2: Suchmaske des ESGF-Knotens am DKRZ.

Weitere Einschränkung der Suche über das Drop-Down-Menü im linken Bereich z.B.:
Schritt 4: Zeige alle Daten-Replikate an (vgl. Abb. 8.3).
Schritt 5: Zeige nur Daten auf den DKRZ Datenknoten an (vgl. Abb. 8.3):
esgf1.dkrz.de: CORDEX + CMIP5 (Daten sind auf Festplatte). Hier sind auch die ReKliEs-De-Daten
esgf2.dkrz.de: CMIP5 Replikate aus dem WDCC (Daten sind auf Band). Hier

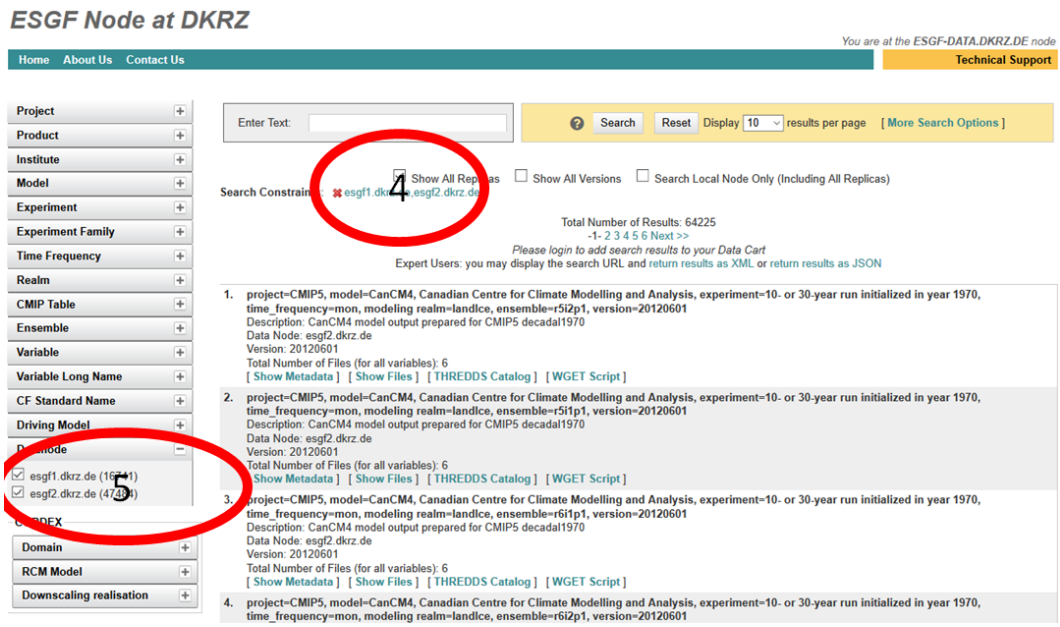

Figure 8.3: Ausgefüllte Suchmaske des ESGF-Knotens am DKRZ. Die rot markierten Zahlen beziehen sich auf die im Text beschriebenen Schritte 4 und 5.

sind nach der Archivierung im WDCC auch die ReKliEs-De-Daten
esgf3.dkrz.de (zukünftig): Deutsche CMIP6 Daten + Replikate
Zur Aktualisierung der Ansicht (mit den entsprechenden Suchkriterien) den „Search"-Button aktivieren. Die Ergebnisliste zeigt die Datensätze, die den ausgewählten Suchkriterien entsprechen.

8.1.1.2 Datensuche und Datendownload (weitere Werkzeuge)

Weitere Werkzeuge für Suche und Download – Es sei darauf hingewiesen, dass die Suchergebnisse als Data Chart gespeichert werden können (wenn eingeloggt). Außerdem ist zu beachten, dass beim Download über den Browser mehrere Dateien gleichzeitig beispielsweise über Downloadmanager heruntergeladen werden können (z.B. DownThemAll unter Mozilla Firefox:

ESGF search API: HTTP-API für die Generierung von Download-Skripten und für die Suche

- https://earthsystemcog.org/projects/cog/esgf_search_restful_api
- Beispiel: Generiere wget Skript für alle Files der Variable cli im Projekt obs4MIPs: https://esgf-data.dkrz.de/esg-search/wget?project=obs4MIPs&variable=cli

OPeNDAP:

- Ermöglicht die Selektion eines Teilgitters und Zeitabschnitts und die Anzeige der zugehörigen Daten in Textform (ASCII CSV) oder

- das direkte Einladen von Daten in Programme wie CDO, ncdump oder R.
- Anleitung: https://portal.enes.org/data/data-metadata-service/search-and-download/opendap

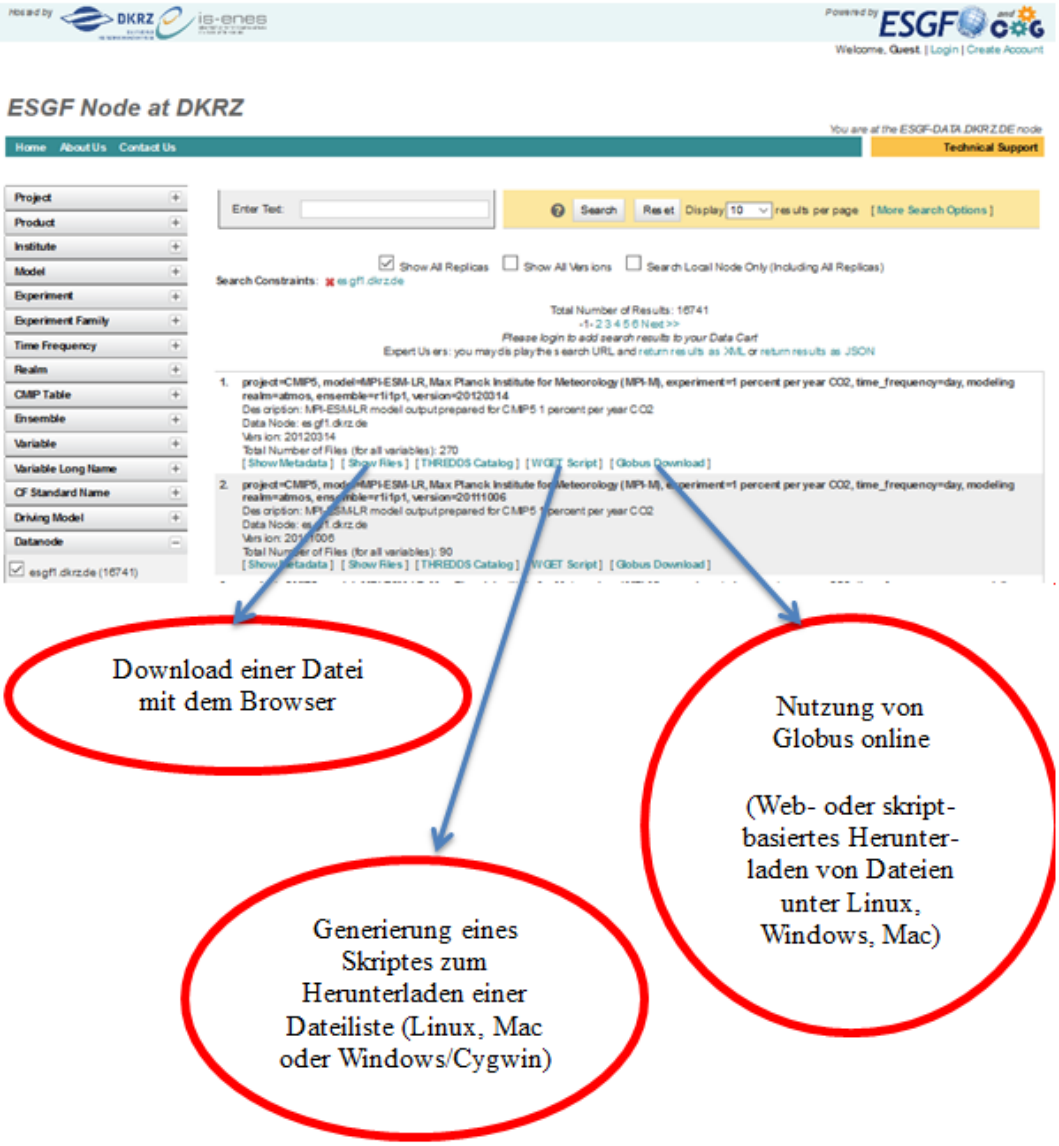

Figure 8.4: Download-Optionen beim ESGF-Knotens am DKRZ.

Globus GridFTP Datenzugriff:

- Voraussetzung: lokale GridFTP Installation
  1. Erzeugung eines Kurzzeitzertifikates:
     `$ myproxy-logon -s esgf-data.dkrz.de -l <user_id> -b -T -t 72 -o ~/.esg/credentials.pem`

2. Setzen von Umgebungsvariablen für den Datentransport:
   $ `export X509_USER_PROXY=~/.esg/credentials.pem && export X509_CERT_DIR=~/.esg/certificates`
3. Datentransfer mit dem Kommando ‚globus-url-copy' auf den lokaler Rechner:
   $ `globus-url-copy -nodcau -dbg gsiftp://esgf1.dkrz.de:2811//cmip5/cmip5/output1/MPIM/MPI-ESM-P/1pctCO2/mon/land/Lmon/r1i1p1/v20120315/c4PftFrac/c4PftFrac_Lmon_MPI-ESM P_1pctCO2_r1i1p1_185001-198912.nc /tmp/test.nc`

esgf-pyclient: ein Python-Client zur Suche im ESGF-Portal

- Für Programmierer
- Nutzt das ESGF Search API
- Installation (Linux): pip install esgf-pyclient
- Quelle ist Github: https://github.com/ESGF/esgf-pyclient
- Dokumentation: https://esgf-pyclient.readthedocs.io/en/latest/
- IPython-Notebook zum Ausprobieren: https://github.com/cehbrecht/demo-notebooks/blob/master/esgf-opendap.ipynb

synda: ein Werkzeug zum Download (und der Synchronisation) von großen ESGF-Datenbeständen zwischen dem ESGF und dem lokalen Rechner

- Nutzt den Python esgf-pyclient zur Datensuche und unterstützt HTTP und GridFTP für den Datendownload
- Quelle ist Github: https://github.com/Prodiguer/synda
- Installation (Linux): source, rpm, docker

Weitere Hilfen zum ESGF sind zu finden unter: User-Tutorials und FAQs: https://www.earthsystemcog.org/projects/cog/tutorials_web
Kontakt:

- Allgemeine Nutzeranfragen an alle ESGF-Betreiber: esgf-user@lists.llnl.gov
- Anfragen zum ESGF beim DKRZ: esgf@dkrz.de
- Anfragen zur eigenen Datenpublikation im ESGF über das DKRZ: esgf-publishing@dkrz.de

- Datendownload aus dem ESGF
Globus per GridFTP (Client)
Probleme: Installation, Zertifikate, Firewall

8.1.2 Datenquelle Langzeitarchiv des DKRZ: WDCC (CERA Datenbank)

8.1.2.1 Datendownload aus dem WDCC per Browser

Eine weitere Quelle für Klimadaten ist das WDCC, das Langzeitarchiv des DKRZ (Abb. 8.5).

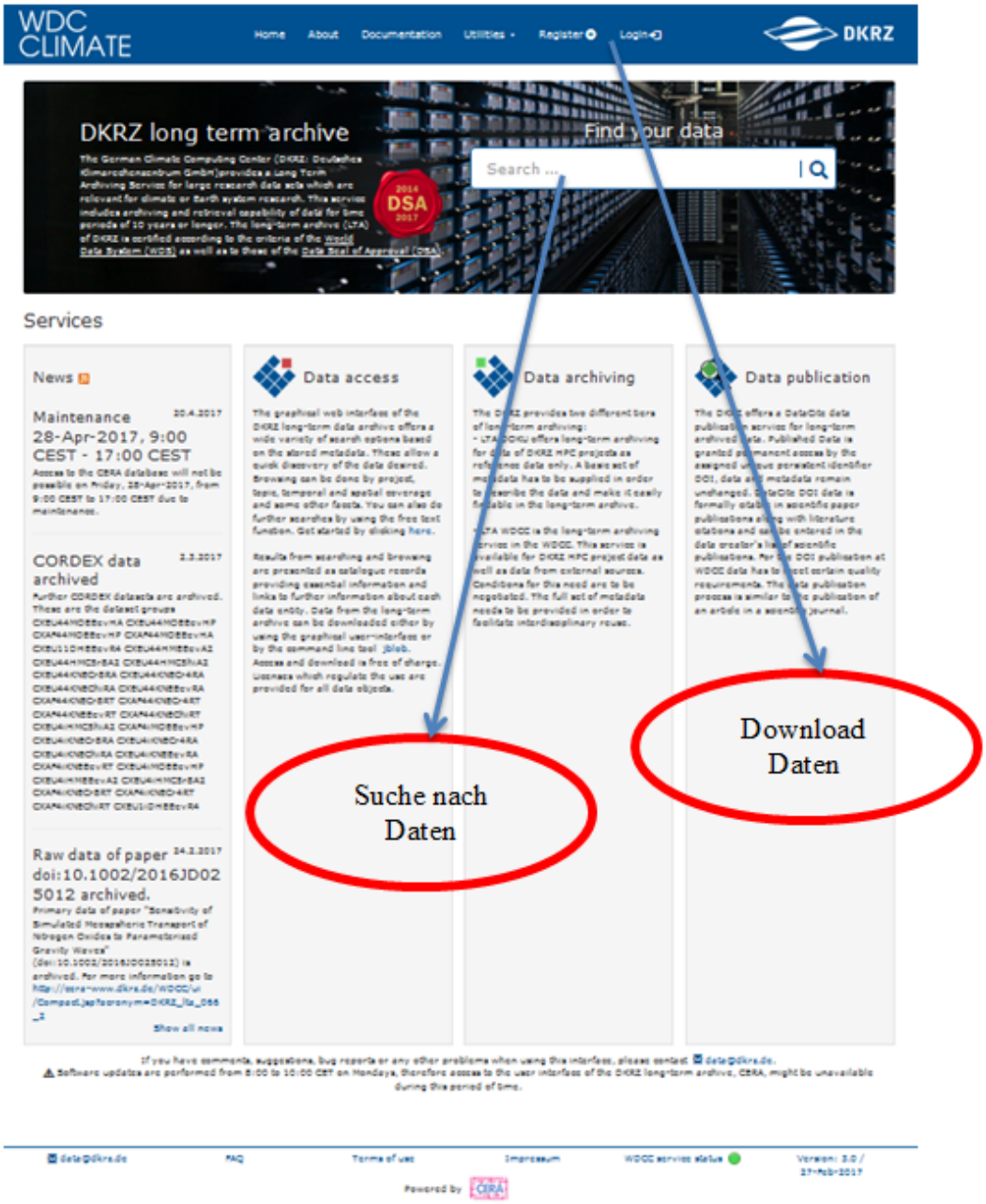

Figure 8.5: Startmaske des WDC Climate am DKRZ

Die Suche in den Metadaten ist frei verfügbar. Sollen Daten heruntergeladen werden, wird ein eigener Zugang benötigt, der in der WDCC-Registrierung

(https://cera-www.dkrz.de/WDCC/ui/cerasearch/register) zu bekommen ist. Benötigt werden eine Mailadresse und ein Institut.

Die Suche nach Daten ist sehr an die Suche nach Daten im ESGF angelehnt. Abb. 8.6 zeigt die Suchoberfläche mit einem Ergebnisbeispiel.

Im Textfeld erfolgt eine Freitextsuche in allen in den Daten vorhandenen Textfeldern.

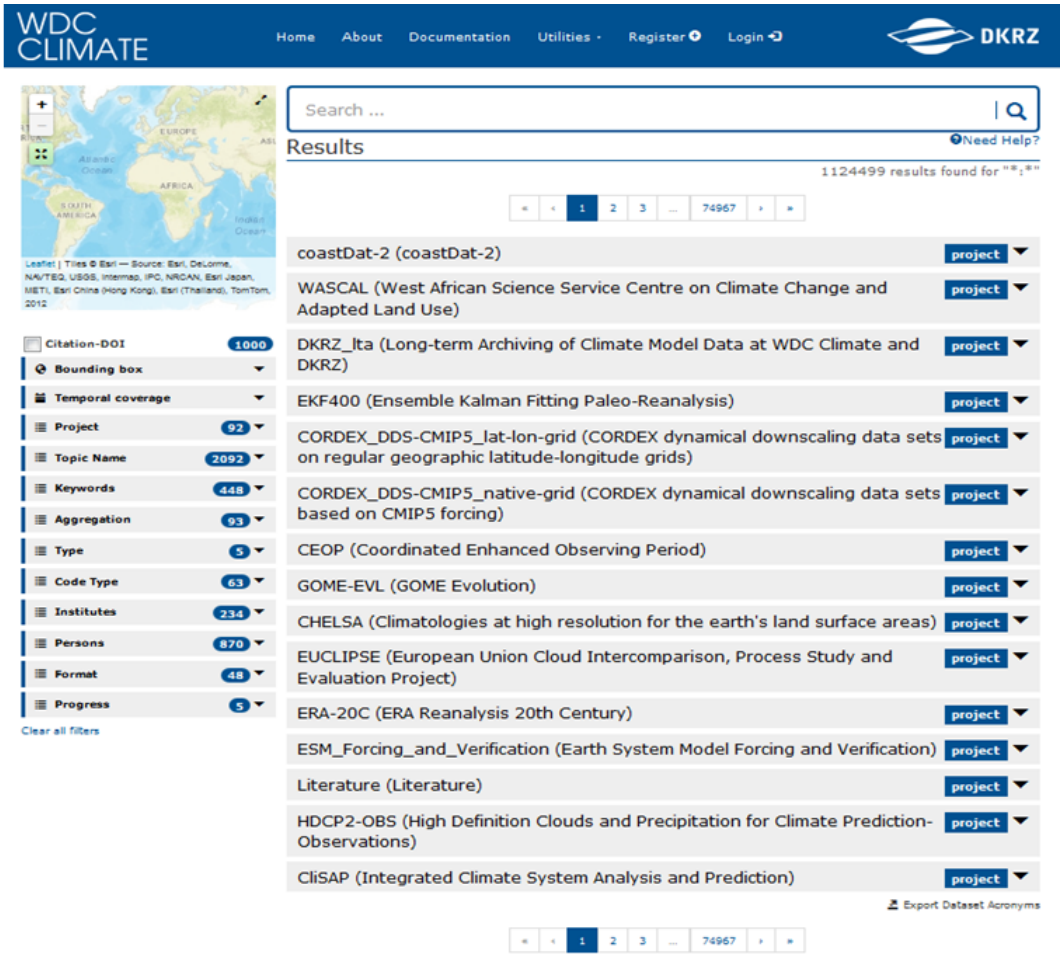

Figure 8.6: Suchoberfläche für Klimadaten beim WDC Climate am DKRZ

Die Auswahllisten auf der linken Seite wiederum ermöglichen eine genauere Suche in den Daten über die verschiedenen Suchkriterien. Als Beispiel ist in Abb. 8.7 die Liste der Projekte gezeigt. Gleichzeitig werden für jede Suchauswahl die Anzahl der vorhandenen Datensätze angezeigt.

Über der jeweiligen Liste können Sortierung und Filterung angegeben werden.

Eine weitere mögliche Einschränkung ist über die Bounding box gegeben (vgl. Abb. 8.8). Es können Zahlenwerte in die Felder eingegeben werden oder es ist eine grafische Auswahl möglich. Suche über Werte (der Button ‚Apply') führt die Suche aus:

Bei der Suche über eine grafische Auswahl aus einer Karte erfolgt die Ausführung wiederum über den Button ‚Apply'. Die Lat/Lon Werte des gewählten Kartenausschnittes werden in den 4 Feldern gleichzeitig angegeben. Wird die Auswahlbox ‚Citation-DOI' markiert (s. Abb. 8.9), werden in der Ergebnisliste nur Datensätze

angezeigt, auf die ein Digitaler Identifier (DOI) vergeben wurde. (Achtung: das kann u.U. eine leere Liste zur Folge haben!)

Abb. 8.10 zeigt die Auswahl eines Experimentes, von dem Daten heruntergeladen werden sollen. Über die Lupe kann die nächste Ebene der Datenhierarchie angezeigt werden (vgl. Abb. 8.11).

In diesem Beispiel gibt es neben den eigentlichen Datensätzen noch ein sog. ‚Additional Info', eine PDF Datei mit weiteren Beschreibungen zu den Daten.

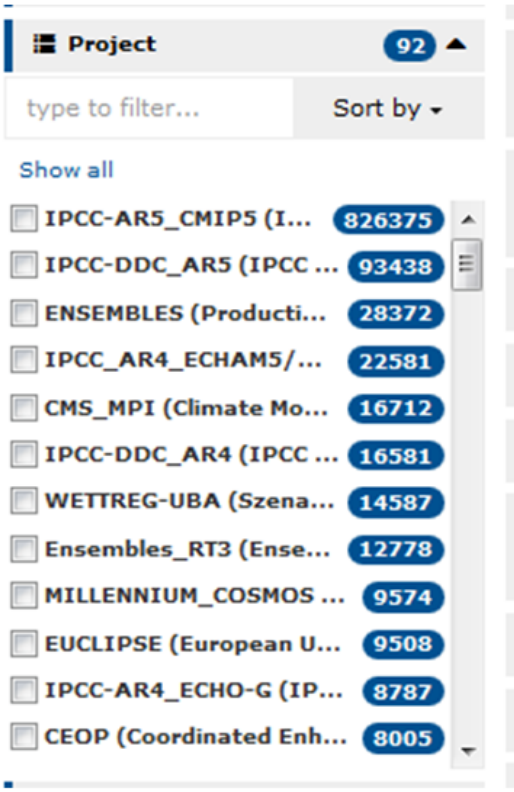

Figure 8.7: Beispiel für Auswahl und Filterung von Klimadaten beim WDC Climate am DKRZ

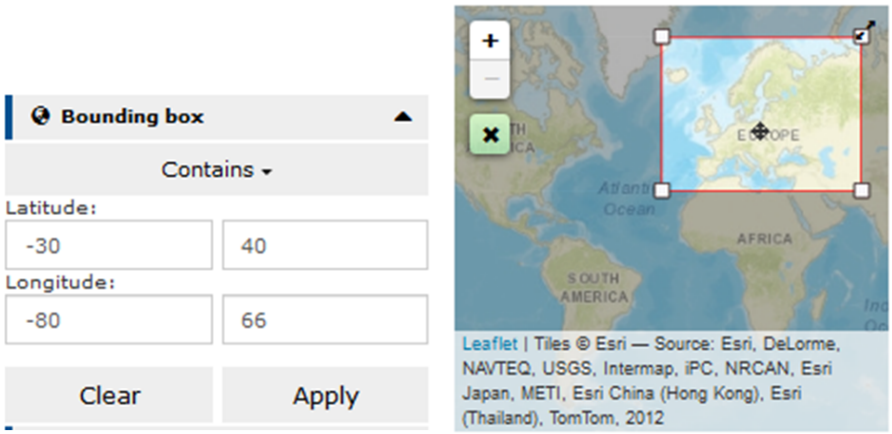

Figure 8.8: Auswahl von Klimadaten aus bestimmten Regionen in Form einer Koordinateneingabe (links) oder eines interaktiven Kartenfelds (rechts).

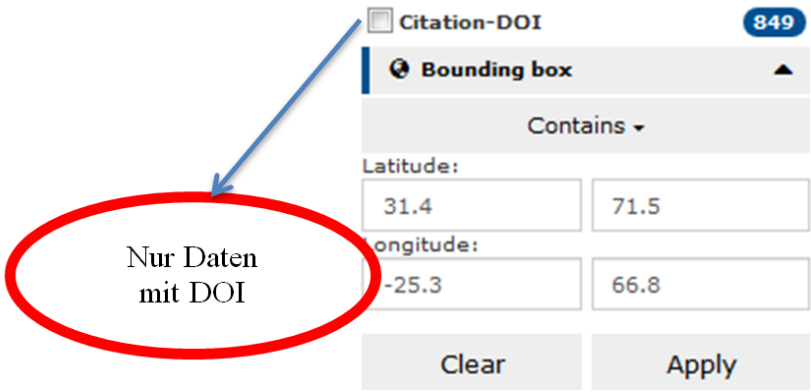

Figure 8.9: Möglichkeit der Einschränkung auf Datensätze mit DOI.

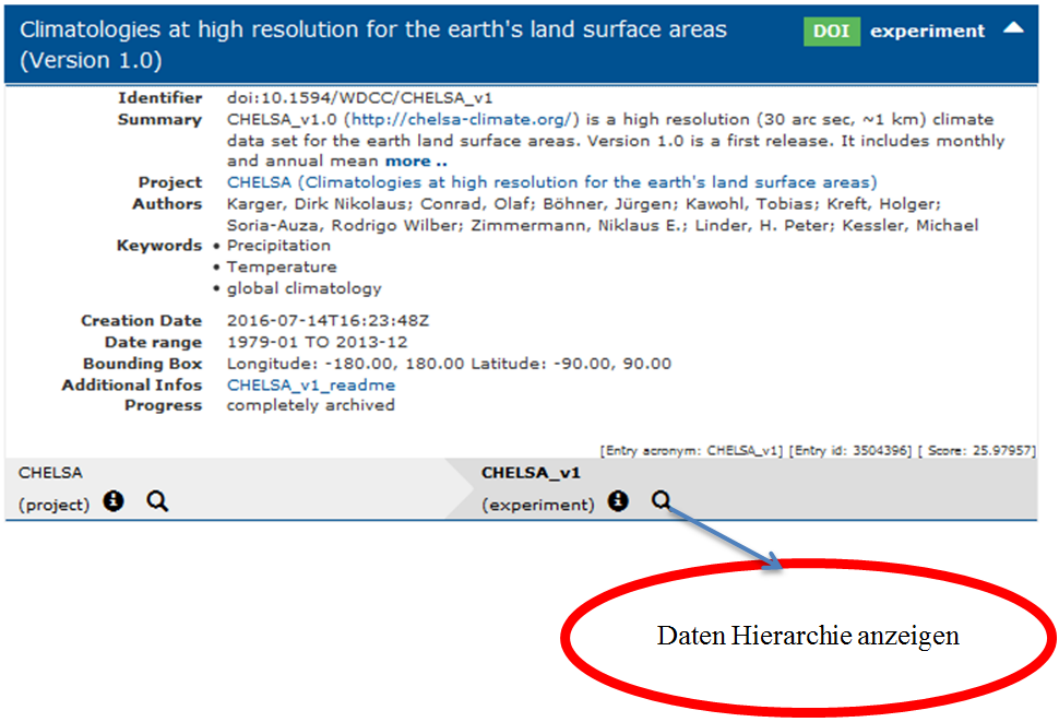

Figure 8.10: Beispiel für die Auswahl eines Experiments.

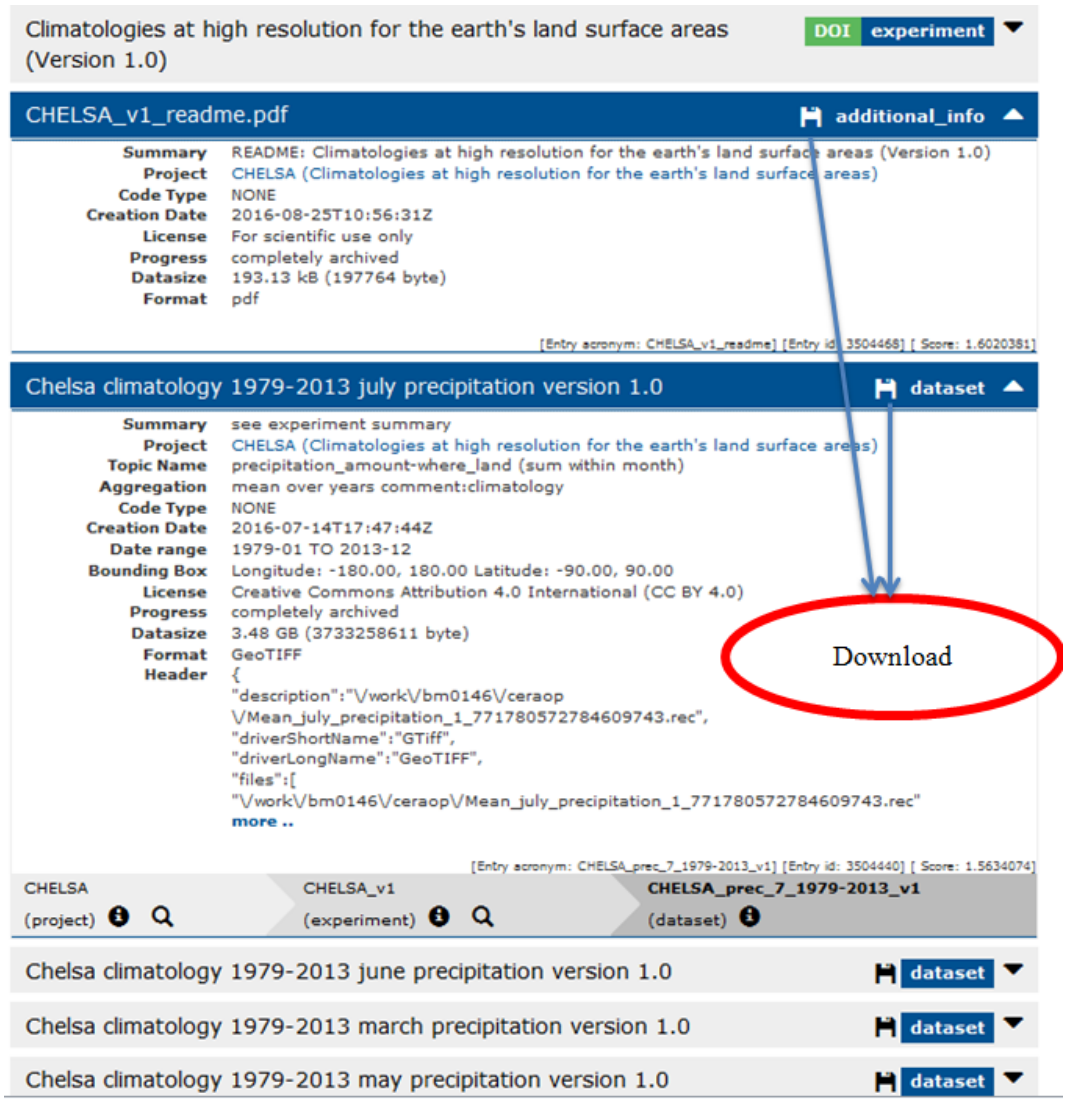

Figure 8.11: Vorbereitung zum Downloads eines ausgewählten Experiments.

Durch klicken auf den Download Link erscheint die Login Seite (Abb. 8.12). Hier werden dann der Benutzername und das Password eingegeben.

Nach erfolgreichem Login können dann die Daten auf den eigenen Rechner geladen werden (Abb. 8.13)

Soll nicht der gesamte Datensatz heruntergeladen werden, gibt es noch die Möglichkeit auf einzelne Perioden (in diesem Fall Monate) einzuschränken. Nach Auswahl ‚Part of dataset' kann ein Zeitbereich gewählt werden. Damit kann die heruntergeladene Datenmenge reduziert werden.

Bei diesem Datendownload direkt aus dem Browser heraus ist es möglich, mehrere vorher gewählte und in den ‚Warenkorb' gelistete Dateien in einem Prozess herunter zu laden.

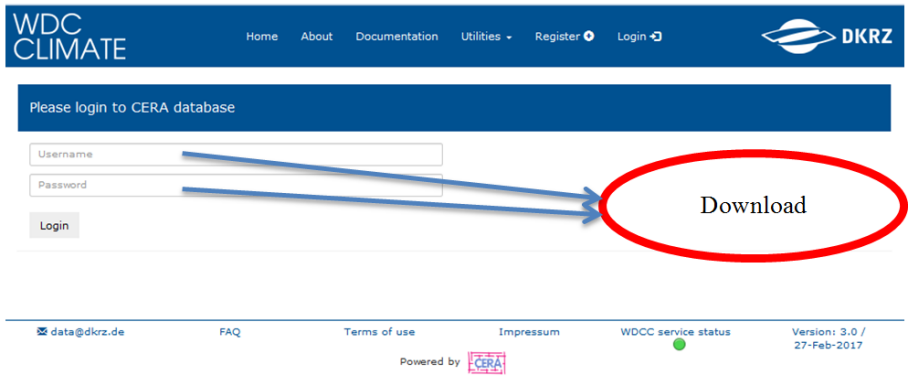

Figure 8.12: Login-Maske der CERA-Datenbank

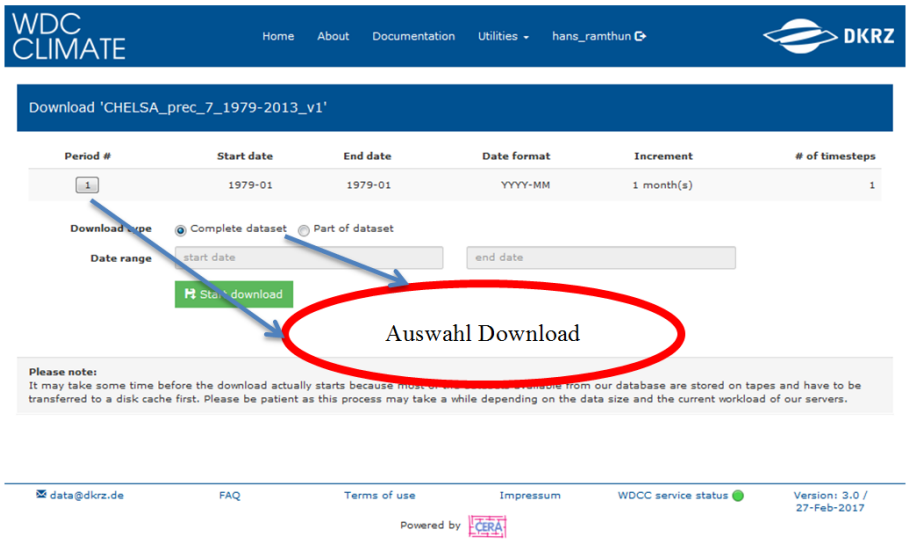

Figure 8.13: Download-Seite der CERA-Datenbank

8.1.2.2 Datendownload aus dem WDCC per Kommandozeile

Eine weitere Möglichkeit Daten aus dem WDCC auf das lokale System zu transferieren ist das Java Kommandozeilen-Programm ‚jblob'. Eine Anleitung zur Installation findet sich hier: https://cera-www.dkrz.de/WDCC/ui/cerasearch/info?site=jblob

Als Voraussetzung ist eine lokale Java Version notwendig.

‚jblob' ermöglich beim Download eine Reduktion der Daten durch die Ausführung von cdo Kommandos vor dem eigentlichen Download. Damit kann die transportierte Datenmenge im Einzelfall deutlich reduziert werden.

Ein Beispiel für das Ausschneiden einer Region (‚Select a specific region of a dataset for download') unter Nutzung der Climate Data Operators (CDO, siehe Abschnitt 8.2.1):

```
jblob --dataset EH5_OM_A1B_1_MM_TSURF --cdo
''sellonlatbox,0,180,-45,45''
```

Die gesamte Dokumentation zum WDCC und der dahinter

liegenden CERA Datenbank ist hier zu finden: https://cera-www.dkrz.de/WDCC/ui/cerasearch/docu

8.1.3 Datendownload aus dem DKRZ Filesystem

Bei Vorhandensein eines DKRZ Zugangs ist es möglich, Daten direkt aus dem DKRZ Filesystem auf den lokalen Rechner zu laden.

Dabei sind verschiedene Möglichkeiten gegeben:

- per ftp (File Transfer Protocol Client)
  Anmerkung: unter Windows kann z.B. das Programm FileZilla benutzt werden
- per Globus GridFTP (Client)
  Anmerkung: das setzt eine vorhandene Globus GridFTP Client Installation voraus mögliche Probleme:
  - Installation muss möglich sein
  - Ein Zertifikat von einer Zertifizierungsstelle muss vorhanden sein
  - Die Firewall muss eine Reihe Ports öffnen (2811 | 2812, 50000–51000)

8.2 Werkzeuge und Methoden zur Verarbeitung der Modellergebnisse

Ein weit verbreitetes Datenformat ist das netCDF-Format, welches von der an der UCAR-Universität beheimateten Gruppe Unidata entwickelt und gepflegt wird. Das netCDF-Format hat sich seit einigen Jahren als das Standardformat für Klimadaten aus allen Bereichen etabliert.

netCDF kann mittlerweile von einer ganzen Reihe von Tools und Programmen direkt, d.h. ohne Formatkonvertierung, verarbeitet werden, beispielsweise durch:

- Kommandozeilen-Werkzeuge, die mit zur netCDF-Software von Unidata gehören: ncdump (Konvertierung nach ASCII) and ncview (einfache Grafik)
- Kommandozeilen-Werkzeuge, z.B. Climate Data Operators (CDO), netCDF Operator (NCO) and UVCDAT: zeige, konvertiere, teile, vereinige und schreibe netCDF-Dateien und führe arithmetische Operationen mit ihnen durch
- Kommandozeilen-Grafik, z.B. NCAR Command Language (NCL) und UVCDAT
- Anwendungsprogramme, z.B. Matlab, R und Ferret
- Python: netcdf4-python ist ein Python-Interface für die netCDF-C-Bibliothek

Kommandozeilen-Werkzeuge haben in der Praxis große Bedeutung. Oft umfassen die heruntergeladenen netCDF-Daten eine zu große Region oder einen zu langen Zeitraum. Um entsprechende Zuschnitte vornehmen zu können, gibt es eine Reihe von Werkzeugen, die aber in der Regel das Betriebssystem Linux beim Nutzer voraussetzen. Zwei von ihnen sollen an dieser Stelle exemplarisch vorgestellt werden.

8.2.1 Datenbearbeitung mit CDO (Climate Data Operators)

Das Max-Planck-Institut für Meteorologie bietet ein kostenloses Programmpaket an, mit dessen Hilfe viele Operationen auf Klimadaten ausgeführt werden können. Es sind u.a. möglich, für Klimadaten:

- Inhalte anzuzeigen
- Zeitliche und räumliche Ausschnitte zu machen
- Daten zu trennen oder zusammenzufügen
- Daten zu ändern, z.B. Umrechnung auf ein anderes Gitter
- ...

Es stehen ca. 500 Operatoren für diese Zwecke in dem Paket zur Verfügung.

Der Download der Programme ist hier zu finden: https://code.mpimet.mpg.de/projects/cdo/files, die Dokumentation und Aufrufbeispiele liegen hier: https://code.mpimet.mpg.de/projects/cdo/wiki

Dieses Programmpaket steht allen Nutzern des DKRZ-HPC-Systems ‚Mistral' in verschiedenen Versionen (1.6.7 ... 1.8.0) über das Laden von entsprechenden Modulen zu Verfügung. Eine Liste kann durch ‚module avail' angezeigt werden, ein Modul wird dann durch ‚module load MODULE' ausgewählt.

Hier ein Beispiel für die Anzeige der Metainformationen einer netCDF Datei – zunächst der Eingabe-Befehl:

```
k204081@mistralpp2% cdo sinfo
/mnt/lustre01/work/kd0956/CORDEX/data/cordex/output/EUR-11/
CLMcom/MPI-M-MPI-ESM-LR/rcp85/r1i1p1/CLMcom-CCLM4-8-17/
v1/day/tasmin/v20140515/tasmin_EUR-11_MPI-M-MPI-ESM-LR_rcp85_r1i1p1_
CLMcom-CCLM4-8-17_v1_day_20060101-20101231.nc
```

Die Ausgabe erfolgt im Eingabefenster. Sie kann optional auch in eine Datei, z.B. im Format ASCII, geschrieben werden. Dafür wäre die Syntax am Ende der Eingabezeile zu ergänzen um „> Asciidatei.txt")

Ausgabe der Datei:

```
File format : NetCDF4 classic ZIP
-1 : Institut Source Steptype Levels Num Points Num Dtype :
Parameter ID
1 : unknown CLMcom-CCLM4-8-17 instant 1 1 174688 1 F32z : -1
Grid coordinates :
1 : lonlat : points=174688 (424x412)
rlon : -28.375 to 18.155 by 0.11 degrees
rlat : -23.375 to 21.835 by 0.110001 degrees
northpole : lon=--162 lat=39.25
Vertical coordinates :
1 : surface : levels=1
Time coordinate : 1826 steps
RefTime = 1949-12-01 00:00:00 Units = days Calendar =
proleptic_gregorian Bounds = true
```

```
YYYY-MM-DD hh:mm:ss YYYY-MM-DD hh:mm:ss YYYY-MM-DD hh:mm:ss YYYY-MM-DD hh:mm:ss
2006-01-01 12:00:00 2006-01-02 12:00:00 2006-01-03 12:00:00 2006-01-04 12:00:00
2006-01-05 12:00:00 2006-01-06 12:00:00 2006-01-07 12:00:00 2006-01-08 12:00:00
2006-01-09 12:00:00 2006-01-10 12:00:00 2006-01-11 12:00:00 2006-01-12 12:00:00
2006-01-13 12:00:00 2006-01-14 12:00:00 2006-01-15 12:00:00 2006-01-16 12:00:00
2006-01-17 12:00:00 2006-01-18 12:00:00 2006-01-19 12:00:00 2006-01-20 12:00:00
2006-01-21 12:00:00 2006-01-22 12:00:00 2006-01-23 12:00:00 2006-01-24 12:00:00
2006-01-25 12:00:00 2006-01-26 12:00:00 2006-01-27 12:00:00 2006-01-28 12:00:00
2006-01-29 12:00:00 2006-01-30 12:00:00 2006-01-31 12:00:00 2006-02-01 12:00:00
2006-02-02 12:00:00 2006-02-03 12:00:00 2006-02-04 12:00:00 2006-02-05 12:00:00
2006-02-06 12:00:00 2006-02-07 12:00:00 2006-02-08 12:00:00 2006-02-09 12:00:00
2006-02-10 12:00:00 2006-02-11 12:00:00 2006-02-12 12:00:00 2006-02-13 12:00:00
2006-02-14 12:00:00 2006-02-15 12:00:00 2006-02-16 12:00:00 2006-02-17 12:00:00
2006-02-18 12:00:00 2006-02-19 12:00:00 2006-02-20 12:00:00 2006-02-21 12:00:00
2006-02-22 12:00:00 2006-02-23 12:00:00 2006-02-24 12:00:00 2006-02-25 12:00:00
2006-02-26 12:00:00 2006-02-27 12:00:00 2006-02-28 12:00:00 2006-03-01 12:00:00

   ...
2010-11-04 12:00:00 2010-11-05 12:00:00 2010-11-06 12:00:00 2010-11-07 12:00:00
2010-11-08 12:00:00 2010-11-09 12:00:00 2010-11-10 12:00:00 2010-11-11 12:00:00
2010-11-12 12:00:00 2010-11-13 12:00:00 2010-11-14 12:00:00 2010-11-15 12:00:00
2010-11-16 12:00:00 2010-11-17 12:00:00 2010-11-18 12:00:00 2010-11-19 12:00:00
2010-11-20 12:00:00 2010-11-21 12:00:00 2010-11-22 12:00:00 2010-11-23 12:00:00
2010-11-24 12:00:00 2010-11-25 12:00:00 2010-11-26 12:00:00 2010-11-27 12:00:00
2010-11-28 12:00:00 2010-11-29 12:00:00 2010-11-30 12:00:00 2010-12-01 12:00:00
2010-12-02 12:00:00 2010-12-03 12:00:00 2010-12-04 12:00:00 2010-12-05 12:00:00
2010-12-06 12:00:00 2010-12-07 12:00:00 2010-12-08 12:00:00 2010-12-09 12:00:00
2010-12-10 12:00:00 2010-12-11 12:00:00 2010-12-12 12:00:00 2010-12-13 12:00:00
2010-12-14 12:00:00 2010-12-15 12:00:00 2010-12-16 12:00:00 2010-12-17 12:00:00
2010-12-18 12:00:00 2010-12-19 12:00:00 2010-12-20 12:00:00 2010-12-21 12:00:00
2010-12-22 12:00:00 2010-12-23 12:00:00 2010-12-24 12:00:00 2010-12-25 12:00:00
2010-12-26 12:00:00 2010-12-27 12:00:00 2010-12-28 12:00:00 2010-12-29 12:00:00
2010-12-30 12:00:00 2010-12-31 12:00:00
cdo sinfo: Processed 1 variable over 1826 timesteps ( 0.16s )
```

8.2.2 Datenbearbeitung mit NCO (netCDF Operators)

Die NCO bestehen zwar nur aus 14 Operatoren, diese sind aber dafür sehr vielseitig und können für die verschiedensten Bearbeitungsaufgaben eingesetzt werden. So kennt der Arithmetik-Operator ncap2 beispielsweise eine lange Liste mathematischer Funktionen und kann fast frei programmiert werden, Schleifen, Verzweigungen und nutzerdefinierte Variablen eingeschlossen. Lesen und Schreiben einzelner Feldkomponenten bis hin zu ganzen mehrdimensionalen Feldern ist möglich. Zur Einbindung in bestehende Programme und Skripte werden Programmbibliotheken für die verschiedensten Skript- und Programmiersprachen angeboten.

NCO-Projektseite mit Links zu Download und Dokumentation: http://nco.sourceforge.net/

Eine detaillierte Darstellung der Möglichkeiten sowohl der CDO als auch der NCO würde den Rahmen dieses Handbuches bei weitem übersteigen. Es sei auf die entsprechende Dokumentation verwiesen.

8.2.3 Datenbearbeitung unter Windows

Ein Problem für viele Nutzer ist die Benutzung der Klimadaten unter dem Betriebssystem Windows. Sowohl CDO und NCO sind zunächst für Linux/UNIX entwickelt worden. Die Installation unter Windows ist schwieriger und die Nutzung nicht selten eingeschränkt.

8.2.3.1 Cygwin

Für Windows-Benutzer bietet sich als Möglichkeit eine Linux-Umgebung zu schaffen, indem das Programm Paket Cygwin installiert wird. Cygwin (https://www.cygwin.com/) gestattet die Ausführung von Linux-Befehlen unter Windows. NCO können unter Cygwin als auch unter Microsoft Visual Studio (MVS) laufen. Es ist aber leider nicht damit getan, CDO oder NCO zu installieren, denn die benötigten Bibliotheken, z.B. netCDF, müssen außerdem auch zur Verfügung stehen. Hier steht der Nutzer vor dem Problem, dass nicht alle benötigte Software auch für alle Windows-Versionen getestet wurde und zur Verfügung steht. Das kann z.B. zur Folge haben, dass z.B. die neu installierten CDO nicht für alle netCDF-Versionen zur Verfügung stehen. Weitere Cygwin Programme oder fehlende Libraries können in dieser Seite gefunden werden: https://cygwin.com/packages/package_list.html. Die Installation erfolgt in Cygwin mit ‚apt-cyg install <NAME_AUS_DER_LISTE>'. Bearbeitungsprogramme wie cdo und nco sind auf den entsprechenden Projektseiten zu finden (vergl. Abschnitte 8.2.1 und 8.2.2). Um Inhalte von netcdf Dateien (z.B. mit ncview) in Cygwin unter Windows anzuzeigen, ist es notwendig in Cygwin eine Umgebungsvariable zu setzen: ‚export DISPLAY=localhost:0.0 ncview' (‚ncview' oder ein anderes Programm welches zur Anzeige benutzt werden soll).

8.2.3.2 Oracle VirtualBox

Eine weitere Möglichkeit Linux-Befehle unter Windows zu benutzen ist die Einrichtung einer „virtuellen Maschine", eines Linux-Rechners, der unter Windows in einer von einem speziellen Windows-Programm geschaffenen Umgebung läuft. Ein Programm, das so eine Umgebung aufbauen und verwalten kann, ist z.B. das kostenlose VirtualBox von Oracle (https://www.virtualbox.org/). Auf diese Weise kann ein „echtes" Linux unter Windows zur Verfügung gestellt werden und die oben beschriebenen Probleme sollten keine Rolle spielen.

8.2.3.3 Docker Container

Das Docker Programm (https://www.docker.com/), welches ebenfalls in Windows installiert werden kann, bietet die Möglichkeit vorgefertigte Docker Container mit allen benötigten Programmen zu starten. Damit ist es ebenfalls möglich, eine sofort benutzbare Linux Umgebung unter Windows zu bekommen.

Sowohl VirtualBox als auch Docker benötigen vorherige Installation auf den lokalen Rechnern und damit auch Administratoren Rechte. Beide Programme sind Freeware und damit kostenlos.

8.2.4 Datenbearbeitung auf der Seite des WDCC

Eine Datenbearbeitung auf der Seite des WDCC kann über verschiedene Wege gemacht werden:

- Jblob CDO-Befehle beim Download
  - Benutzung der CDO-Operatoren sollte bekannt sein
- WPS (Web Processing Service) am DKRZ
  - Datenzugriff ist nicht für alle Daten eingerichtet
  - noch keine Produktionsversion
  - WPS am DKRZ: https://bovec.dkrz.de/
- Download, Datenreduzierung und Bereitstellung in der DKRZ-Cloud ist auf Anfrage in Ausnahmefällen möglich, dies als automatisierter Service ist erst für die Zukunft geplant, der Zugriff erfolgt über ein Webinterface: https://swiftbrowser.dkrz.de/login/

8.2.5 Datenbearbeitung beim Endanwender

Für die Bearbeitung von Daten, die beim Endanwender liegen, sind verschiedenen Werkzeuge benutzbar:

- Beschreibung von netCDF-Tools auf der DKRZ-Seite
  - https://www.dkrz.de/Nutzerportal-en/doku/data-processing/tools
- Bearbeitung unter Linux
  - CDO (siehe: https://code.mpimet.mpg.de/projects/cdo/)
    Region: `cdo sellonlatbox,10,70,-35,23 input-file output-file`
  - NetCDF Programme (ncdump, ncview, ncgen, nccopy)
    http://www.unidata.ucar.edu/software/netcdf/workshops/2011/utilities/index.html
  - Python netcdf4 Library
    http://unidata.github.io/netcdf4-python/

- Bearbeitung unter Windows
  - CDO (nicht die aktuellste Version)
    https://code.zmaw.de/projects/cdo/wiki/Win32
    Für die Linux Umgebung in Windows ‚Cygwin' sind auch neuere cdo Versionen benutzbar
  - Python-netCDF4-Library
    Problem: Installation aller Komponenten
    http://unidata.github.io/netcdf4-python/

8.2.6 Datenvisualisierung beim Endanwender

Für die Visualisierung von Daten beim Endanwender ist das Programm Panoply zu empfehlen.

Das Programm Panoply ist unter Windows, MacOS und Linux lauffähig und ist hier zu bekommen: https://www.giss.nasa.gov/tools/panoply/ Eine Installation ist nicht notwendig. Für die Benutzung ist eine JAVA 8 Umgebung notwendig.

Die Visualisierung kann mit lokalen Dateien oder mit Dateien, die über eine URL zu erreichenden sind, erfolgen.

Eine Anleitung zu Panoply in Deutsch ist hier zu finden: http://bildungsserver.hamburg.de/contentblob/4490584/ a5e00eae6aa9beec8f90b90434ed21af/data/kurzanleitung-panoply-4.pdf

Verschiedene Beispiele für cdo und nco Aufrufen sind in Abschnitt 8.4 zu finden.

8.3 Umgang mit Datenlücken

8.3.1 Modelldaten

Rechnungen von GCMs (Globale Klima Modelle) und RCMs (Regionale Klima Modelle) werden lückenfrei durchgeführt. Bei einigen Modellen, z.B. bei der HadGEM-Familie, enden diese für das 21. Jahrhundert bereits im Herbst 2098. Die Berechnungen von Größen wie Mittelwerten für Dekaden erfolgen in diesem Fall auf unvollständiger Basis. Nicht vorhandene Werte in den Dateien werden mit ‚MissingValue' oder ‚FillValue' Werten aufgefüllt. ‚MissingValue' oder ‚FillValue' sind Zahlenwerte die außerhalb des möglichen Wertebereiches der Daten liegen und damit nicht vorhandene Daten sicher markieren.

8.3.2 Beobachtungsdaten

Insbesondere die Datenreihen an Messstationen können Lücken enthalten. Eine Möglichkeit damit umzugehen ist, eine lückenbehaftete Reihe komplett auszuschließen. Alternativ sollte auf Nutzerseite geklärt werden, ob die zu ermittelnden Größen auch mit Lücken für die jeweilige Nutzer-Anwendung zu benutzen sind oder mit welchen Einschränkungen diese Nutzung erfolgt. Weiterhin

ist zu klären ob fehlende Messwerte aufgefüllt werden sollen. In diesem Fall sollte aber ein Kriterium erstellt werden, wie groß eine Lücke maximal sein darf und wie viele Daten insgesamt fehlen dürfen, damit eine Auffüllung noch sinnvoll ist und die Reihe nicht verworfen werden muss. Verbindliche Kriterien existieren nicht – in der Aufbereitung für IDP (Auswerte Programm der Bundesländer für Klimadaten) wurden beispielsweise Reihen zur Fehlwertersetzung akzeptiert, bei denen 10% Fehlwerte enthalten sind. Es ist außerdem zu bedenken, dass bei einer Niederschlagszeitreihe ein relativ großes Risiko besteht, dass an fehlenden Tagen auch Niederschlagsereignisse stattgefunden haben. Größen wie Monatssummen können daher zu gering sein. Andere Größen, wie Temperatur oder Luftdruck sind gegenüber Datenlücken nicht so sensitiv.

Kleinere Lücken (in täglichen Zeitreihen bis zu einer Länge von 7 Tagen) werden in der einfachsten Variante durch Verfahren aufgefüllt, die mit Hilfe der Werte vor und nach der Lücke eine lineare Interpolation vornehmen. Etwas aufwendigere Verfahren bilden einen mittleren Jahresgang der Größe und rechnen die Daten in Anomalien zu diesem Jahresgang um. In die Interpolation fließt dann neben dem angenommenen linearen Verlauf im Fehlzeitraum auch klimatologische Information ein.

Es gibt außerdem die Möglichkeit, fehlende Werte für einen Zeitpunkt aus den Daten der Nachbarstationen zu konstruieren. Solche Verfahren können komplex entwickelt sein, da durchaus das Risiko besteht, dass Nachbarstationen ebenfalls Lücken besitzen und ein Kriterium entwickelt werden muss, wie weit eine Station entfernt sein darf, damit sie noch hinzugezogen wird. Die Lückenfüllung kann in einem solchen Fall mit einer wechselnden Gruppe von Stationen erfolgen. Eine weitere Abwägung umfasst die Zahl der Nachbarstationen, aus denen die Fehlwertersetzung berechnet werden soll.

Sollten Fehlwertersetzungen erfolgen, so ist zu dokumentieren, nach welchen Kriterien, mit welchem Verfahren und in welchem Umfang dies geschah.

8.4 Beispiele für Datenbearbeitung

8.4.1 CDO Kommandos

Umgebungsvariablen (unter Linux) für CDO:
```
export IGNORE_ATT_COORDINATES=1
export CDO_HISTORY_INFO=0
```

1. Selektion der Monate 3...11 aus einer Datei:
   `cdo -f nc4c -seasmean -selmonth,3/11 input_file output_file`

2. Selektion des Monats 12 aus einer Datei:
   `cdo -f nc4c -seasmean -selmonth,12 input_file output_file`

3. Selektion von Zeitschritten aus einer Datei:
   `cdo -f nc4c -s timselmean,3,2 input_file output_file`

4. Zusammenbinden zweier Dateien (z.B. Winde u + v Variablen)
   `cdo merge Input1_file Input2_file Output_file_u_v`

5. Entrotieren von Variablen u + v (Winde):
   `cdo rotuvb,u,v Output_file_u_v Output_file_entrotiert`

6. Selektion einer Variablen aus einer Datei:
   Für u: `cdo selvar,u Output_file_entrotiert Output_file_entrotiert _Var_u`
   Für v: `cdo selvar,v Output_file_entrotiert Output_file_entrotiert _Var_v`

7. Selektion eines Jahres aus einer Datei:
   `cdo -f nc4c -s selyear,1981 input_file output_file`

8. Verschiebung der Zeit in einer Datei um 30 Tage:
   `cdo -f nc4c -s -shifttime,30days input_file output_file`

9. Setzen einer neuen Zeitmarke:
   `cdo -f nc4c -s setreftime,'days since 1950-01-01 00:00:00' input_file output_file`

10. Operatoren sind kombinierbar: Abarbeitung von rechts nach links!
    `cdo -f nc4c -s selyear,1997 -shifttime,27days -setreftime, 1950-01-01, 00:00:00 input_file output_file`

11. Selektion eines Levels 3 (z.B. aus einer Datei mit verschiedenen Schichten einer Bodenvariablen):
    `cdo -f nc4c sellevidx,3 input_file output_fil`

12. Multiplikation der Variablen mit einem Faktor:
    `cdo mulc,0,02745 input_file output_file`

13. Selektion einer bestimmten Stunde (oder mehrerer Stunden) aus einer Datei:
    `cdo -f %s -s selhour,2,3,4 input_file output_file`

14. Erstellung von Monatsdaten aus täglichen Daten:
    `cdo -f nc4c -s -monmean -daymean .mulc,27,3%s input_file output_file`

15. Die Zeitachse neu definieren (incl. time_bnds):
    `cdo --cmor setreftime,1980-11-10,10:00:00,seconds infile outfile`

8.4.2 NCO Kommandos

1. Umwandeln einer Variablen von Typ ‚double' (altes Zahlenformat) in Typ ‚float':
   `ncap2 -s 'MyVar=float(MyVar)' input_file output_file`

2. Multiplikation der Werte einer Variablen mit einem Faktor:
   `ncap2 -O -s 'MyVar=MyVar*0.001' input_file output_file`

3. Druck von Pa nach hPa umrechnen:

   - sehr langsam
     `ncap2 -s 'pressure=pressure/100.' wa12clmN_eraint-ctrl_cclmprs_2011-12.nc prs_1.nc`
   - sehr schnell
     `ncatted -O -a units,pressure,o,c,'hPa' prs_1.nc`

4. Variable umbenennen:
   `ncrename -h -O -v Altername,Neuername Datei.nc`

5. Setzen eines Attributes (z.B. tracking_id):

   - mehrere Attribute gleichzeitig ändern/hinzufügen:
   - `$VAR` = Variablenname
   - `$UNIT` = Einheit
   - `SNAME` = standard_name
   - `uuid` erstellen (unter Unix): `uuid='echo $(python -c 'import uuid; print uuid.uuid1()')'`
   - `LNAME` = long_name
   -
     ```
     ncatted -h -O -a _FillValue,$VAR,o,f,1.e20
     -a missing_value,$VAR,o,f,1.e20
     -a units,$VAR,o,c,$UNIT
     -a standard_name,$VAR,o,c,$SNAME
     -a tracking_id,global,o,c,''$uuid''
     -a long_name,$VAR,o,c,''$LNAME'' Datei.nc
     ```

6. Lat / Lon Variablen aus einer Datei in eine Zwischendatei speichern:
   `ncks -v lat,lon pr_mon_19790101-19800101.nc latlon.nc`
   `ncap2 -s 'rlat=double(rlat); rlon=double(rlon)' latlon.nc dlatlon.nc`
   - und Einbau in eine vorhandene Datei:
     `ncks -A dlatlon.nc MeineDatei.nc`

8.4.3 Jblob Version 2 Dokumentation (nur auf Englisch verfügbar)

8.4.3.1 Download

Important: Please read the Terms of use document available here!

Click here to download Jblob V2.0.10

If you have Jblob 2.0.5 or 2.0.9 already installed you can download the updated file here, just replace the old file in the installation directory with the downloaded one.

8.4.3.2 Installation instructions

Jblob on DKRZ's mistral system

Jblob is already installed on mistral and available as a module, just execute 'module load jblob'.

Prerequisites

Jblob needs an installed Java Runtime Environment (either JRE or JDK of version 6 or above).

Please note: GCJ is NOT supported! If you install a Sun JDK/JRE in addition to a pre-installed gcj-based java distribution make sure to change the global JAVA_HOME environment variable to the new setting.

You can download the Java Runtime Environment at http://www.oracle.com/technetwork/java/javase/downloads/index.html.

The Installation steps are

1. Download the program archive linked above. This file contains all files neccessary for use with Windows and Unix.

2. Extract the contents of the archive to the desired installation directory.

3. Unix/Linux: Edit the lines containing 'JBLOB_HOME' and 'JAVA_HOME' in the jblob script.

Windows: Edit the lines containing 'JBLOB_HOME' and 'JAVA_HOME' in the jblob.bat batch script.

Changes for use with HTTP-Proxies: This is only necessary if your internet connection requires use of a proxy!

The java options `-Dhttp.proxyHost` and `-Dhttp.proxyPort` can be used to enable jblob connections over http-proxies.

For unix/linux systems add the following options to the last line of the jblob script (before the 'de.dkrz.cera.applications.JblobClient $" ' part)

`-Dhttp.proxyHost=<your_proxy_host> -Dhttp.proxyPort=<your_proxy_port>`

For windows systems add the following options to the last line of the jblob.bat batch script (before the 'de.dkrz.cera.applications.JblobClient %*' part

`-Dhttp.proxyHost=<your_proxy_host> -Dhttp.proxyPort=<your_proxy_port>`

8.4.3.3 Documentation

Synopsis

```
jblob --dataset | -d dataset_acronym [ options ]
jblob --cmip5file drs_name [ options ]
jblob --help
jblob --version
```

Options

–cdo cdo operator CDO processing during data download. See section "CDO processing" for details.
–config file Read username and password information from file instead of ∼/.netrc.
–dateext Append timestamp(s) to filename instead of record number(s). Works only for datasets with a "calendrical" temporal structure.
–dir directory The transferred dataset will be placed in the specified directory.
–file filename The transferred dataset will be given the specified name instead of datasetname_rmin-rmax.
–origin LTA filename Transfer a single file from a dataset containing DOKU/LTA data.
–password password Use password for database login. Works only in conjunction with –username.
–quiet Quiet operation. No output except error messages.
–rmax number Last record to transfer. If omitted, the transfer will end with the last available record.
–rmin number First record to transfer. If omitted, the transfer will start with the first available record.
–tmax date Last date to transfer. The standard date format is yyyy-mm-ddThh:mm, however unnecessary parts may be omitted (i.e. yyyy-mm is sufficient when working with monthly data).
–tmin date First date to transfer. The standard date format is yyyy-mm-ddThh:mm, however unnecessary parts may be omitted (i.e. yyyy-mm is sufficient when working with monthly data).
–username username Use username for database login. Works only in conjuction with –password.

Supplying username and password

By default the program tries to read the username and password information from the ∼/.netrc file. To utilize this you must add an entry for machine 'cera' to your .netrc file (i.e. `machine cera login your_username password your_password`). Alternatively you can specify a configuration file from where the login information will be read by using the –config option. This configuration file must follow the same syntax as the .netrc file. As a last resort you can provide username and

password with the –username and –password options.

Exit values

0 Success
1 Unspecified error Please check program output for additional messages (i.e. do not use the -quiet option).
3 Invalid dataset name The requested dataset name is doesn't exist in the database.
4 Invalid dataset id The requested dataset id doesn't exist in the database.
5 Insufficient permission to dataset You do not have the neccessary permission to access the requested dataset. Please contact for further assistance.
8 I/O error An I/O error was encountered during the data transfer.
12 Temporal structure not supported for tmin/tmax selection The selected dataset cannot be downloaded by selecting start and end dates, please use selection by record numbers (–rmax, –rmin) instead.
13 Selected start/end record/time outside of datasets coverage. The supplied rmin/rmax/tmin/tmax value(s) are outside of the dataset's coverage.
14 Dataset is empty (data_size = 0) Please contact for further information.

CDO Processing

The following CDO (1.6.9) operators are currently available for use with Jblob:

```
sellonlatbox --cdo ''sellonlatbox,<min_lon>,<max_lon>,<min_lat>,
                  <max_lat>''
selcode --cdo ''selcode,<code>''
selgridname --cdo ''selgridname,<grid_name>''
```

CDO processing is usable with these dataset formats: GRIB, NetCDF, IEG

Please note: processing may take some time, please be patient and DO NOT start the same download again if the first one doesn't start downloading immediately!

Examples

a) Transfer all records of dataset EH4OPYC_SRES_A2_WIND10 to /tmp, reading username and password from ∼/.netrc

```
jblob --dataset EH4OPYC_SRES_A2_WIND10 --dir /tmp
```

b) Transfer data from Jan. 2001 to Dec. 2010 of dataset EH5_OM_A1B_1_MM_TSURF, reading username and password from command line

```
jblob --dataset EH5_OM_A1B_1_MM_TSURF --tmin 2001-01 --tmax 2010-12
        --username foo --password bar
```

c) Download a single file from a LTA dataset

```
jblob --dataset DKRZ_LTA_564_ds00001 --origin
''DMS-paper/dms-paper-all-data-for-fig.tar''
```

d) Download CMIP5 data by drs_name

```
jblob --cmip5file ''cmip5/output1/MPI-M/MPI-ESM-MR/amip/day/
        atmos/day/r1i1p1/v20120330/clt/clt_day_MPI-ESM-MR
        _amip_r1i1p1_19790101-20081231.nc''
```

e) Select a specific region of a dataset for download

```
jblob --dataset EH5_OM_A1B_1_MM_TSURF
        --cdo ''sellonlatbox,0,180,-45,45''
```

Known Issues

Contents of DOKU/LTA datasets currently can only be downloaded from DKRZ HPSS tape archive separately due to large file sizes.