

## Zwischenbericht April 2017

Zuwendungsempfänger: <i>Deutsches Klimarechenzentrum GmbH</i>	Förderkennzeichen: <i>01LP1605A</i>
Vorhabenbezeichnung: <i>CMIP6-DICAD Teilprojekt 1</i>	Titel von TP1: „Bereitstellung des nationalen Beitrags zur Datenbasis des IPCC/AR6 sowie Unterstützung der CMIP6*-Aktivitäten in Deutschland: Aufbau und Betrieb des Nationalen CMIP6-Datenarchivs und einer Infrastruktur zur Datenqualitätssicherung
Laufzeit des Vorhabens: <i>1. Juli 2016 – 30. Juni 2020</i>	
Berichtszeitraum: <i>1. Juli 2016 – 30. April 2017</i>	

### Vorbemerkungen

Alle geförderten Stellen konnten zwischenzeitlich besetzt werden, allerdings mit einer zeitlichen Verzögerung von 2,5 Monaten bei zwei der vier Stellen.

Am 19. 7. 2016 fand am DKRZ ein Projekt-Kickoff-Treffen beider beteiligter Verbünde statt. Am 24.1. das erste DICAD-interne Projekttreffen, in dem sich Projektpartner über ihre Fortschritte und Pläne ausgetauscht haben. Dazu wurden auch Kollegen vom MPI-M eingeladen, die nicht Projektpartner sind, aber natürlich in die Projektaktivitäten involviert sind. Im Oktober 2017 werden im zeitlichen Zusammenhang mit dem 2-jährigen DKRZ-USER-Treffen auch Schulungen über die im Vorhaben entwickelten Werkzeuge durchgeführt.

Zur besseren Koordination der Partner im Projekt wurde ein Redmine (<https://redmine.dkrz.de/projects/cmip6-dicad>) aufgesetzt, in dem jeweils die Fortschritte der einzelnen Arbeitspakete (AP) und relevante Dokumente zu finden sind.

Die vom BMBF gewünschte ‚repräsentative‘ Projekt-Webseite wurde aufgesetzt (<https://www.dkrz.de/c6de>).

### Wichtige wissenschaftlich-technische und andere Ergebnisse

#### AP2 (CDOs)

- Ein Prototyp des CDO-Operators zur CMIP-konformen Formatierung von Modellergebnissen steht zur Verfügung. Dieser Operator benutzt noch die CMIP5-Version der CMOR-Bibliothek (CMOR2), da deren Entwicklung zum Beginn der Projektarbeit abgeschlossen war. Der Prototyp wurde für sämtliche mit MPI-ESM1 berechneten CMIP5-Variablen, die eine Teilmenge der CMIP6-Variablen sind, erfolgreich getestet.
- Der Operator erfüllt auch spezielle Anforderungen, wie beispielsweise Erzeugung von Archivvariablen mit Koordinatenachsen vom Typ ‚Charakter‘, die zur Bündelung von Variablen benutzt werden (z.B. Vegetationstypen). Solche Datenstrukturen sind bis jetzt mit dem Datenmodell der CDOs unvereinbar. Mit der Einführung dieser Datenstrukturen hat der Einstieg in die Anpassung des CDO-Datenmodells an das CMIP6-Datenmodell stattgefunden.
- Für andere Variablen, die vollständig auf Modellkoordinaten abgegeben werden sollen, muss wegen der hybriden Sigma-Druck-Koordinaten Zusatzinformation an die CMOR-Bibliothek übergeben werden. Die angestrebte Modellunabhängigkeit des Operators macht das Programm sehr komplex, da keine Annahmen über die zu verarbeitenden Daten gemacht werden können.
- Ein großer Teil der Anpassung des CDO-Metadatenmodells an das in CMIP definierte ist erfolgt.
- Eine erste Version eines Nutzerhandbuchs steht zur Verfügung (<https://redmine.dkrz.de/documents/639>).
- Gegenwärtig wird die Anbindung an die CMIP6-Version von CMOR (CMOR3) vorgenommen und für denselben Variablenumfang getestet.
- Der CDO-Operator sollte in der Lage sein, beide Standards (CMIP5 und CMIP6) zu bedienen, da vermutlich auch in Zukunft noch (z.B. CORDEX-)Daten in Anlehnung an CMIP5/CMOR2 formatiert werden. Zurzeit ist die Planung einen einzigen CDO-Operator zu entwickeln, der nur an die CMOR3-Bibliothek angebunden wird, aber auch CMIP5/CMOR2-kompatible Daten erzeugen kann. Gegebenenfalls müssen zwei Operatoren (cdo cmor2,... und cdo cmor3) für die jeweiligen Anwendungen zur Verfügung gestellt werden.

#### AP3 (Anpassung an Standard)

- Die Variablenliste der Datenanforderung hängt von den MIPs, an denen teilgenommen werden soll, sowie vom Experiment ab. Für die Erzeugung so einer Variablenliste wurde ein python-basiertes Softwaresystem CMIP/WIP<sup>1</sup>-seitig zur Verfügung gestellt. Um eine vereinfachte Anwendung der komplexen Software zu erreichen, wurde im Vorhaben eine webbasierte graphische Nutzeroberfläche (GUI) entwickelt.
- Wenn Variablen im Modelloutput nicht dieselben Namen haben wie die Projektvariablen – und damit muss gerechnet werden, muss die Zuordnung der Variablen dem CDO-Operator bekannt gemacht werden. Diese Zuordnungsliste, ‚Mapping-Table‘ genannt, kann sehr umfangreich werden (bis zu ca. 2100 Zeilen). Auch für die Erzeugung von Mapping-Tabellen wurde ein webbasiertes GUI entwickelt.
- Der in AP2 entwickelte CDO-Operator wurde in das für CMIP5 benutzte Skript zur Formatierung der MPI-ESM1-Ergebnisse eingefügt.
- Die Parallelität der Verarbeitung im Skript wurde erhöht, um der geplanten Verwendung einer größeren Anzahl von Prozessoren bei der Modellintegration und der höheren Auflösung Rechnung zu tragen. Damit bleibt das für eine reibungsfreie operationelle Verarbeitung wichtige Verhältnis der Wallclock-Zeit von Integration und Datenverarbeitung (ca. 6/1) erhalten.
- Drei potentielle Datenverarbeitungsmodule (Aggregation, Diagnostik und Formatierung mit CMOR3) wurden separiert, um eine vereinfachte Anpassung der Skripte an andere ESM-Modelle zu erreichen.
- Die Vorbereitung der Skripte für EMAC ist in Kooperation mit den DLR-Partnern in Arbeit.
- An einer automatisierten Erzeugung des 3. Moduls der Datenaufbereitung (Formatierung mit CMOR3) entsprechend der projekt- und experimentabhängigen Variablenlisten wird gegenwärtig gearbeitet. Die automatisierte Erzeugung dieses Moduls) basiert auf den ESM-abhängigen Mapping-Tabellen.
- Gleichzeitig mit den Angaben zum Mapping der Modell- und CMIP-Variablen können auch Angaben zu eventuell noch benötigten Diagnostiken in das GUI eingegeben werden. Auf diesen Angaben wird eine automatische Erzeugung des 2. Moduls des Datenverarbeitungsskripts basieren.

#### AP4 (Qualitätsprüfung)

- Dem am DKRZ entwickelten Programmpaket zur Qualitätssicherung (QA-DKRZ) wurde ein Modul mit CMIP6-kompatiblen Strukturen hinzugefügt. Darüber hinaus wurde ein Nachbearbeitungsschritt eingeführt, der Annotationen, d.h. Auffälligkeiten bezüglich der Konventionen, aus den Logdateien extrahiert und in einem Dokument im JSON-Format zusammenfasst.
- Auf der „6<sup>th</sup> Annual ESGF F2F“ Konferenz in Washington, D.C., USA, (Report: <http://esgf.llnl.gov/reports.html>) wurde vereinbart,

<sup>1</sup> CMIP-Working Group on Coupled Modeling Infrastructure Panel

dass CMIP6-Daten eine Minimal-QA erfüllen sollen, um in den ESGF-Datenpool aufgenommen zu werden (näher erläutert ist dies in dem Dokument „CMIP6 ESGF Publication Requirements“,

[https://docs.google.com/document/d/1oFie37HV0cCpKrcOEvc\\_ATTAoeklm2zc3ZEfzCdWC4M/edit](https://docs.google.com/document/d/1oFie37HV0cCpKrcOEvc_ATTAoeklm2zc3ZEfzCdWC4M/edit));

1. die CMIP6 eigene „Data Reference Syntax“ (DRS) von Pfad und Dateinamen ist korrekt,
  2. die Metadaten genügen der CF Konvention v1.6,
  3. die Metadaten müssen dem CMIP(6)-Standard-Validator „PrePARE“, der von D. Nadeau am „Lawrence Livermore National Laboratory (PCMDI)“ entwickelt wird, genügen;
- Die Punkte 1 und 2 werden vom QA-DKRZ Programm verifiziert, Punkt 3 durch PrePARE.
  - PrePARE wurde als externes Programm in die QA-DKRZ-Skripte eingebunden. Der Arbeitsablauf ist wie folgt:
    1. ein Teilbereich der CMIP6-Daten wird dateiweise selektiert und durchlaufen,
    2. das DRS-Modul des QA-DKRZ testet die DRS-Struktur für Pfad und Dateinamen,
    3. PrePARE testet die CMIP(6)-Kompatibilität und liefert Annotationen in wenig strukturierter Textform,
    4. der PrePARE-Output wird in QA-DKRZ konforme Annotationen umgewandelt und zusammen mit eventuellen DRS-Annotationen in einer Logdatei gespeichert.
  - Eine erste QA-DKRZ-Version wurde auf GitHub zur Verfügung gestellt.
  - CMIP6 betreffende Einträge wurden der QA-DKRZ-Dokumentation hinzugefügt.
  - Die QA-DKRZ-Dokumentation wird allgemein auf <https://readthedocs.org/projects/qa-dkrz> bereitgestellt.
  - Ein Prototyp des web-basierten QA-DKRZ-Dienstes für einzelne Projektdateien („Spot Check“) wurde eingerichtet.

#### **AP5 (CMIP6 Datenarchiv)**

- Die Infrastruktur eines nationalen CMIP6<sup>+</sup>-Datenknotens wurde beim DKRZ eingerichtet. Erste Beobachtungsdaten der DLR wurden hinzugefügt, aber noch nicht publiziert.
- Auf der ESGF-F2F-Konferenz wurde eine „Testföderation“ vereinbart, um vorab die Infrastruktur für die Replikation von CMIP6-Daten zu errichten und zu testen. Teilnehmende Institute: CEDA (UK), DKRZ, IPSL (F), NCI (AUS) und PCMDI (USA). Stand:
  1. Installation des Replikationswerkzeugs „synda“, das am IPSL entwickelt wird;
  2. Testumgebung für den Index- und Datenknoten wurde in die Testföderation integriert;
  3. pseudo-CMIP6 Daten, d.h. CMIP5-Datensätze (500GB) mit den korrekten CMIP6-Metadaten (DRS, Filenamen, ...) wurden publiziert und mittels http bzw. gridFtp repliziert;
  4. auf eine Authentifikation wurde zunächst verzichtet, da diese in der Testföderation zunächst Probleme bereitet;
  5. das automatisierte Publizieren von Daten durch synda wurde eingerichtet und getestet,

#### **AP7 (Zitierfähigkeit im CMIP6+-Archiv)**

- Das Zitationsmodul für die Frühzitation ist mit der Basisfunktionalität für Nutzer ist freigeschaltet.
- Das Metadatenschema wurde entwickelt, ein Import unter Verwendung des CMIP6-Vokabulars ist geschrieben. Anpassungen aufgrund von Änderungen in der Struktur des CMIP6-Vokabulars werden kontinuierlich durchgeführt.
- Eine graphische Nutzeroberfläche basierend auf Oracle Application Express (APEX) wurde mit der Basisfunktionalität entwickelt und hat erste Nutzer. Ein Tutorial zur Nutzung der Oberfläche wurde erstellt. Erweiterungen für eine verbesserte Nutzerfreundlichkeit und Skalierbarkeit der Anwendung stehen noch aus.
- Eine API-Schnittstelle für den Maschinenzugriff auf Zitatinhalt wurde unter Verwendung von ‚content negotiation#‘ umgesetzt, die ‚Landing Page‘ für die Zitatlinks ist erstellt.
- Die Anpassungen im ESGF-Publisher sowie den Portalen (CoG) wurden in Kooperation mit den Kollegen vom PCMDI innerhalb des ESGF-QCWT (Quality Control Working Team) durchgeführt. Diese nutzen die API-Schnittstelle des Zitatsservices.
- Innerhalb des internationalen CMIP6-Projekts wurden ferner Diskussionen zur Einbettung der Citation in CMIP6 mit ES-DOC und dem WIP geführt.
- Der CMIP-Datenpool wurde als neuer IPCC/DDC-Service mit den IPCC Working Group TSU<sup>2</sup>s sowie innerhalb von TGICA diskutiert. Eine genaue Spezifikation von Ablauf und Service stehen noch aus.
- Zur Vorbereitung der Langzeitarchivierung wurde eine Registrierung von erweiterten Metadaten (ancillary metadata) im ESGF im Rahmen der ESGF-QCWT implementiert.
- Input4MIPs (CMIP6 externe Daten) wurde als Anwendungsfall für die Frühzitation ergänzt, die Daten repliziert und dem DICAD-Konsortium zur Verfügung gestellt.
- Die Implementierung von Funktionalitäten zur DataCite DOI-Registrierung und zur Verteilung der Zitatinformationen über den OAI-Server außerhalb des Projektkontextes wurde begonnen.

#### **AP8 (Langzeitarchivierung)**

- Schwerpunkt der Arbeit ist derzeit die Analyse des CMIP6-Metadatenmodells und seine Abbildung auf das Modell des Langzeitarchivs. Das CMIP6-Metadatenmodell liegt noch nicht final vor. Für die bereits abgestimmten Elemente wird ein erster Entwurf für den ‚Crosswalk‘ beider Modelle entwickelt.
- Dafür ist die beidseitige projektspezifische Definition der Semantik erforderlich.
- Die Schnittstelle zwischen dem sich noch in Erprobung befindlichen „CMIP6 Early Citation – Service“ und dem Langzeitarchiv wird untersucht.

## **Vergleich mit der Arbeits-, Zeit- und Ausgabenplanung**

Im **AP1 (CMIP6<sup>+</sup>-Experimente)** sind im Berichtszeitraum die Meilensteine M1-M3 fällig:

*M1(Mo 3) - DECK-Experimente mit MPI-ESM1 gestartet*

*M2(Mo 6) - historical Experimente mit MPI-ESM1 gestartet*

*M3(Mo 9) - Szenarien-Experimente mit MPI-ESM1 gestartet.*

In Bezug auf die Durchführung der DECK-Experimente mit MP-ESM1 hat es eine Änderung insofern gegeben, dass diese Experimente nicht, wie im Antrag geschrieben, am DKRZ durchgeführt werden, sondern am MPI-M. Dies geschah auf Wunsch des MPI-Ms. Die Experimente von ScenarioMIP sollen aber weiterhin unter Federführung vom DKRZ durchgeführt werden.

Da die Entwicklung der CMIP6-Version von MPI-ESM1, insbesondere in Bezug auf die im Modell geplanten Diagnostiken für die angeforderten Variablen, noch nicht abgeschlossen ist, konnten weder die DECK- und historical Experimente noch die Szenarien-

<sup>2</sup> Technical Support Unit

Experimente gestartet werden. Keine der internationalen Modellierungsgruppen hat bisher die CMIP6-Experimente gestartet, da die Finalisierung der angefragten Variablen noch aussteht. Der Prozess ist zeitaufwändiger als geplant wegen der Vielzahl der endorsed MIPs in CMIP6. Die Entwicklung befindet sich aber weiterhin im IPCC-AR6 Zeitplan.

Im **AP2 (CDOs)** war nur ein Meilenstein fällig:

*M1(Mo 9) - CDO-Formatierungsoperator für CMIP5-Variablen nutzbar.*  
Dieser Meilenstein wurde erreicht.

Zu **AP3 (Anpassung an Standard)** ist zu bemerken, dass es eine Verzögerung von 2,5 Monaten bei der Besetzung der Stelle gab, die sich mit AP3 beschäftigt. Die Auswirkungen werden dadurch abgefedert, dass weder die Experimente am MPI-ESM1 gestartet wurden, noch die Software für den CMIP6-Datenrequest fehlerfrei und in finaler Form vorliegt. Die fälligen Meilensteine sind

*M1(Mo 3) - Skript für Nachbereitung der MPI-ESM1-Daten*

*M2(Mo 6) - Werkzeug zum Prüfen des CMIP-Variablenumfang*

*M3(Mo 9) - Skript zur Formatierung von CMIP6-Daten.*

M1 konnte nicht ganz erreicht werden, da die Modelldiagnostik noch nicht abgeschlossen ist, und deshalb keine Testdaten zur Verfügung stehen. M2 ist erreicht, und das erstellte Werkzeug wird benutzt, um eine automatisierte Erzeugung von Variablenlisten zu ermöglichen, die den geplanten Aktivitäten angepasst sind. In Bezug auf M2 hat sich also die Komplexität der Aufgabe deutlich erhöht.

M3 wird wie oben erwähnt in Form der an den Variablenumfang angepassten automatischen Erzeugung des Skriptes gerade bearbeitet.

Im **AP4 (Qualitätsprüfung)** sind im Berichtszeitraum die Meilensteine M1 – M3 vorgesehen:

*M1(Mo 3) - Release der ersten CMIP6<sup>+</sup> angepassten Version des Werkzeugs zur Qualitätssicherung.*

Dieser Meilenstein wurde erreicht.

*M2(Mo 3) - Dokumentation zur Installation und Nutzung des Werkzeugs zur Prüfung von CMIP6<sup>+</sup>-Daten (erste Version, wird basierend auf Erfahrungen der Endnutzer verbessert).*

Dieser Meilenstein wurde nur teilweise erreicht; es wurde lediglich ein Prototyp der Dokumentation erstellt, da Erfahrungen von Endnutzern aufgrund fehlender Daten nicht vorliegen.

*M3(Mo 6) - Bereitstellung des Spot-Check-Dienstes im Internet.*

Dieser Meilenstein wurde erreicht.

#### **AP5 (CMIP6 Datenarchiv)**

*M1(Mo 6) – Bericht zur operativen Installation eines ESGF Datenknotens für CMIP6 am DKRZ mit Anbindung an das nationale CMIP Datenarchiv.*

Dieser Meilenstein wurde insofern erreicht, als dass ein Testknoten eingerichtet und die Publikation/Replikation in der ESGF erprobt wurde.

#### **AP7 (Zitierfähigkeit im CMIP6<sup>+</sup>-Archiv)**

Die Frühzitation ist auch Teil des internationalen CMIP6-Projekts. Damit ist ihr Zeitplan abhängig vom Zeitplan des internationalen Projekts in Bezug auf die Frühzitation, der eingehalten ist.

Zum fälligen Meilenstein:

*AP7 M1(Mo 6) - Festlegung des zu verwendenden Datenbankmodells und Anbindung an die Schnittstelle der Datenbank von Errata und Annotationen*

Das Datenbankmodell für die Frühzitation ist festgelegt und wird operativ verwendet. Schnittstellen der CMIP6-Dienste ‚Errata und Annotationen‘ (siehe AP4) sind derzeit noch nicht verfügbar, so dass die Übernahme dieser Informationen in die Metadaten des Langzeitarchivs (Datendokumentation) noch nicht geplant werden kann.

Die Stelle, die sich mit **AP8 (Langzeitarchivierung)** beschäftigt, konnte nur mit 2,5 Monate Verspätung besetzt werden. Durch Einsatz von DKRZ-Mitarbeitern wurden jedoch die ersten Arbeiten begonnen, so dass sich insgesamt keine Verzögerung bei der Erreichung der Projektziele ergibt. In dem Berichtszeitraum gibt es in diesem AP keine Milestones.

Die einzige Abweichung von der **Ausgabenplanung** war bedingt durch die verzögerte Einstellung zweier Projektmitarbeiter.

## **Geänderte Aussichten für die Erreichung der Ziele des Vorhabens**

Keine

## **Neue relevante Ergebnisse von dritter Seite**

### **AP3**

Zur Zeit der Antragsschreibung war nicht klar, dass die CMIP6-Variablenliste derart umfangreich (ca. 2100) und die Abhängigkeit von den CMIP6-endorsedMIPs so komplex sein würde. Zur Erzeugung ‚individueller‘ Variablenlisten wurde von CMIP6-Projektseite eine Software zur Verfügung gestellt. Allerdings sind wir der Meinung, dass diese Software nicht sehr benutzerfreundlich ist. Deshalb wurde beschlossen, ein benutzerfreundliches GUI zu entwickeln, das auch für automatisierte Erzeugung von ‚individuellen‘ (d.h. Aktivität und Modellen angepassten) Datenverarbeitungsskripten verwendet wird.

## **Änderungen in der Zielsetzung**

Keine

## **Fortschreibung des Verwertungsplans**

Keine Änderung