CMIP6 DICAD

Lessons Learned - Sammlung

Hierunter verstehen wir in erster Linie Feedback bzgl. des Vorgehens im DICAD-Projekt und weniger Feedback bzgl. der technischen Organisation von CMIP6 allgemein (Datenrequest, Antriebsdaten, Verzögerungen, Datenstandard, ...). Dieses Thema können wir in der allgemeinen Diskussion (s. u. im Dokument) kurz anreißen.

HPC-Ressourcenplanung

- **Gemeinschaftsanträge in Zukunft modellbezogen in separaten RZ-Projekten.** Bei gemeinsamen Anträgen kommt es vor, dass ein Modell dem anderen Ressourcen blockiert. Hier empfehlen wir zwar weiterhin wenn möglich eine DKRZ-nahe administrative Leitung für alle Modelle, aber eine Aufteilung auf mehrere Rechenzeitprojekte (Vorbild: Konstrukt von PalMod). Genehmigte Ressourcen können bei Bedarf weiterhin kurzfristig auf Nachfrage anderen Projektpartnern zugewiesen werden.
- Templates für große Vergleichsprojekte mit Standard-ESMs wären hierbei hilfreich.
- **Platz auf /work wird unterschätzt.**
    - Ursprünglich wurden Ressourcen für zu rechnende Experimente beantragt. Da kamen über die Zeit hinzu:
        - Auf /work gehaltene Experimente für Standardisierung
        - Der standardisierte Output

      Verbesserung durch Beschleunigung
        - des Standardisierungs- und Publikationsworkflows
        - der Archivierung und Rearchivierung
        - Kennzahlen von Antrag 0 potenzieren sich
- **Rechenzeit und Priorisierung** kann am DKRZ ausnahmsweise kurzfristig und kurzzeitig flexibel vergeben werden (Gott sei Dank)
    - Kurzfristig und kurzzeitig konnten uns über 500 TB Speicher von anderen Projekten zugewiesen werden
    - Ebenfalls wurde uns auf Mistral für die hochaufgelösten ScenarioMIP-Simulationen mit AWI-CM und MPI-ESM Priorität eingeräumt, sodass wir alle Simulationen innerhalb von ca. zwei Monaten abschließen konnten.
- Datenvolumenberechnung für Pool mittels DReq ist eine sehr grobe Abschätzung (s. Vortrag AP3).

Modellrechnungen

- Einiges an Rechenaufwand bzw. Post-Processing und errata-Fälle hätte man sich durch bessere Kommunikation ersparen können.
- Dezentrale Organisation erschwert den Support von manchen MIPs
    - Start von HighResMIP (Primavera), MiKlip, DCPP an den Instituten zu unterschiedlichen, zum Teil sehr frühen Zeitpunkten die andere Modellausgaben nutzen
    - Anpassung des Post-processing an den einzelner MIPs schwierig oder nicht umgesetzt
    - Am besten wird in Zusammenarbeit eine Namelist aufgestellt, sodass die Anforderungen einer jeden Gruppe (bzw. MIPs) erfüllt werden können und nur selten die Nachberechnung von Diagnostiken oder Variablen nötig wird.
-

Post-Processing-Infrastruktur (cdo cmor, WebGUI)

- **Modellierer sind während der Hauptentwicklung der PP-Infrastruktur mit Modellentwicklung beschäftigt, was eine User-nahe Softwareentwicklung erschwert**
    - Wünsche und Kritik werden nicht bzw. zu spät eingebracht
    - Umsetzung eines anwenderfreundlichen GUIs wurde in einem Fall nicht angenommen; stattdessen wurde ein eigenes Programm für FESOM geschrieben. Dessen Anwendung ist per Skript einfacher, aber vermutlich weniger flexibel. In jedem Fall ein weiteres CMORlight
    - Lösungsansatz: Die CMIP DECK Experimente schneller, automatischer und operationeller durchführen, in dem Modellentwicklung an einem Punkt abbrechen muss - mit genug Zeit zur Anpassung von Output und Schnittstellen
        - Kommentar: Man will sein Modell aber auch nicht unter Wert verkaufen und eventuelle Bugs etc. beseitigen. Zudem müssen um den Datenrequest erfüllen zu können neue Diagnostiken in das Modell eingebaut bzw. die Modellausgabe angepasst werden. Die Entwicklung des DReqs bzw. das Sammeln der Datenanforderungen sowie die Erstellung der Forcingdaten fand parallel zu alledem statt.
- **Schnittstellen für die PP-Infrastruktur spielen bei der Modellentwicklung eine untergeordnete Rolle**
    - Testdaten sind nicht oder zu spät verfügbar.
    - Lösung?

- PP-Infrastruktur ist modellunabhängig und sehr modular aufgebaut und muss in Zusammenarbeit mit den Modellierern in das modelleigene Scripting integriert werden
    - Im “Standalone”-Betrieb redundante und somit nutzerunfreundliche PP-Konfigurationen
    - Ansprechpartner für jedes Modell nötig, der von Anfang die Entwicklung der Infrastruktur verfolgt, Feedback gibt, und sich für die Integration in das modelleigene Skripting engagiert
- **Idee I: Modularität ermöglicht flexible Anwendung über Projektende hinaus.**
    - Wird umgesetzt
- **Zur Datenstandardisierung fehlt Wissenschaftlern**
    - **, die Rohdaten auswerten *können*, ein Anreiz.** Paper und Konferenzbeiträge mit Datenanalysen haben aus verschiedenen Gründen Priorität gegenüber Möglichmachung der Datennachnutzung.
    - **Zeit**
      So werden HiWis kurzfristig und kurz zur Standardisierung angestellt. Diese
        - Sind nicht da bei offiziellen Workshops und müssen separat geschult werden
        - Können in kurzer Zeit nicht den Scope von CMIP6 erfassen/verstehen

      Andere Ansätze: Qualitätslevel statt Qualitätsstandard (Vorbild: AtMoDat). Umkehrung der Methodik: Statt Fehlersuche Qualitätsbeweis.
- **Workshops zum Capacity buildung brauchen mindestens 2 Tage**
    - Der CMIP Datenproduktionsworkflow berührt sehr viele WPs
    - Vorschlag für die Zukunft: Regelmäßiger, operationeller Workshop über eine Woche mit allen Projektpartnern (remote?)

Publikationsworkflow (QA, Ticketsystem, ESGF, Citation)

- Integration des Vollständigkeitschecks der Zitatsinformationen in den Workflow (Das ist innerhalb von CMIP6 eine offizielle Aufgabe der ESGF Data Node Manager.)
- Publikation eines kompletten Experiments manchmal nicht in einem Schwung
    - Kommt das bei anderen Knoten vor, kann eine vollständige Replikation nicht garantiert werden
    - Doppelte Variablenpublikation weil von 2 Submodellen in 2 Schwüngen geliefert
    - → Stärkere Formulierung/Durchsetzung von Publikationsregeln, Einbau von Tests bei Datenabgabe in die QA

- Tickets und QA für einen eindeutig definierten Datenworkflow viel zu zeitintensiv
    - → Zukünftig weniger händisch, mehr operationell und automatisch

Auswertung (ESMValTool, Freva CMIP6 Datenportal)

- Großer Fokus von CMIP6-Auswertenden auf python basierte Infrastruktur wie die von Pangeo
    - Intake-esm erlaubt Selektion mit Hilfe von Katalogen und dem Paket panda und ist eine Alternative für Freva.
- Feedback Loops nicht ideal. Versionierung von Ergebnis-Bildern, die bestehen bleiben. Warum wurden Bilder entfernt? Wurden daraufhin Daten ausgetauscht?

Organisation und Zusammenarbeit

- Ausschuss für Fragen in Zusammenhang mit ESMValTool/Monitoring, gewählt beim Kickoff, daraufhin nie wieder in Erscheinung getreten. Trotz der Verzögerungen auf Seiten der CMIP6-Datenanforderung und damit erst spät feststehenden Modell-Namelists hätte diese Arbeitsgruppe etwas bewirken können. Da jedes Modellsystem aber bereits ein Monitoringmodul besitzt, fällt das nicht zu sehr ins Gewicht. Eine Grundlage für die Übernahme des ESMValTool-Quicklook-Monitorings auch für AWI-CM und MPI-ESM wurde zumindest geschaffen.
- Github statt svn, google documents statt docx hin und herschicken, regelmäßigere Meetings (insbesondere rein technische Meetings für Feedback zur PP-Infrastruktur, Modelldiagnostiken etc).
-

Lessons Learned - GoToMeeting-Diskussion

Tido: Organisation gut. Bottleneck: CMORisierung. Inkonsistenzen zwischen CMOR-Software und QA/prepare durch häufige Änderungen des CMIP-Variablenstandards. Für CMIP7 keine späte Änderung, v.a. im DReq. Einfrieren wenn das erste Modell anfängt zu rechnen. Versch. DReq Versionen an versch. Datenknoten laufen dem Ziel von CMIP, einen einfachen Modellvergleich durch einen gemeinsamen Datenstandard/Variablenstandard ermöglichen, zuwider. CMORisierung für FESOM hat gut funktioniert, verschiedene Skriptfragmente erschweren die CMORisierung von ECHAM. Ziel am DKRZ war die flexible Anwendung auf versch. Modelle. Viele Fehler in die man reingelaufen ist, die nur durch Support gelöst werden konnten.

Bernadette: CMORisierung betrifft auch andere Projekte. Viel Arbeit ist entstanden, für die es möglicherweise schon Lösungen am DKRZ gibt. Grundsätzliche Überarbeitung des Konzepts CMORisierung. Nachlieferung von Daten auf Grund von Verzögerungen durch CMORisierung nicht umsetzbar. Wie geht man mit dem running target / schrittweise Anpassung um? CMORisierungsaufwand wird unterschätzt. Unterschiede im Datenstandard über verschiedene Projekte. Längere timeslots für CMORisierung vorsehen. Verständnis des Prozesses fehlt.
Es gibt zwei Möglichkeiten: a) Der Datenrequest ist nicht final, aber das Modell. Hier müssten dann nach den Modellrechnungen aufwändig Anpassungen bzw. Arbeiten stattfinden um fehlenden Output nachzuliefern bzw. vorhandenen Output zu standardisieren. b) Der Datenrequest ist final, nun müssen teils zeitintensive Anpassungen am Modelloutput bzw. den Modelldiagnostiken vorgenommen werden bevor die Rechnungen starten können.
Vergl. PRIMAVERA. Verfahren für unterschiedliche Datenrequests?
Frühe DReq Versionen konnten nicht veröffentlicht werden, HighResMIP als Teil von CMIP und PRIMAVERA (Anm. Fabi). Es gab hier also Unterschiede zwischen den Voraussetzungen von CMIP und den Voraussetzungen für eine Publikation im DKRZ-ESGF-Knoten.
Michael: Unterschiede in DRS bei Primavera waren auch in den HighResMIP-Daten zu finden. Dadurch konnten die Daten nicht als CMIP-Daten VÖ werden.
Michael: Was ist die Erwartungshaltung der Modellierenden bzgl. CMORisierung? Festlegung des Standards? Änderungen abwärtskompatibel? Kompletter Prozess zu kompliziert?
Tido: Internationale Kollaboration, EC-Earth macht eigene CMORisierung. Ziel: Einheitlicheres Tool.
Michael: CMOR ist ein einheitliches Tool. Verlangt viel Arbeit. Ziel der Post-Processing-Infrastruktur in DICAD war es, diese Arbeit zu erleichtern und für die deutschen Modelle nachhaltig zu unterstützen.
Bernadette: Modellierer und Datenleute bewegen sich auseinander, man versteht die Sicht der anderen nicht. Datenstandard so kompliziert, dass Modellierer dem nicht mehr Nachkommen können.
Michael: Was können die Modellierungsgruppen investieren an Arbeit? Ist CMORisierung Teil der Modellierung oder DM?
Christian: Aus CORDEX Sicht: Immer schon aufwendig und es wird aufwendiger. “Notwendiges Übel”. Hauptproblem der Standardisierung ist das Mapping. Das wird immer gebraucht und zeitliche Aggregation erfordern am meisten Arbeit. Das Hinzufügen der globalen Attribute ist dagegen weniger aufwändig. Braucht Know-How von beiden Seiten, Modellierung und DM.
Michael: NIcht klar, wer die Arbeit macht.
Christian: Die Modellierer geben die Daten ab, sind in der Verantwortung. Auch in ReKlies wurde die Arbeit unterschätzt. Nur mit DKRZ-Hilfe machbar (Frank).
Björn: Sich ändernder DReq hat auch Probleme beim ESMValTool hervorgerufen

(repräsentativ für alle nachgeordneten Nutzer der ESGF-Daten). Wichtig wäre ein eingefrorener DReq.
Martin: Datenanforderungen wurden von jedem MIP separat erstellt und an zentraler Stelle gesammelt. Am Anfang hatte man so manche Variable doppelt und dreifach im Datenrequest. Die Konsolidierung dauerte sehr lange. Hier müsste in Zukunft aus einer vorhandenen Liste von Variablen der jew. DReq der MIPs ausgewählt werden um diese Konsolidierung überflüssig zu machen.

Christopher: Flexibles Handeln im durchgeplanten Projekt muss möglich sein. Positive Zusammenarbeit mit DLR.
Was schief geht kommt nicht wieder bei den Entwicklern an. Feedbackloops passen nicht.
Operationelles System: Erfordert viel Arbeit und bedarf Förderung.
Fabian: Gibt es genug Möglichkeiten zum Feedback geben? Braucht es einen Workshop?
Christopher: Passive Rückmeldung würde reichen, braucht kein Zugriff auf die Daten.
Lisa: Genug Werbung für ESMVal in der Community.
Björn: 2 Problemklassen: Wiss. Diskussion einerseits, technische Problem andererseits. Technische Probleme werden lieber “offline” diskutiert.
Christopher: Versionierungsinfos für die erzeugten Plots, Versionierungsinfos / Errata-Informationen der Daten durchreichen/weiterreichen.
Björn: Versionierungsinfos der Plots sind in den Provenance-Infos in den Metadaten verfügbar und müssten “nur” verknüpft werden.
Christopher: Idee best. Plots/Datensätzen zu folgen und bei Änderungen per Mail benachrichtigt zu werden. (+1 Björn)

Christian: 2 Teile des APs. ICON-Entwicklung Hilfe durch versch. Personen. Treffen zur Strategieausrichtung zw. DWD, MPI-M geplant. Leitlinien für CMIP aufzustellen wäre evtl. etwas für die Agenda.
Johann: CMIP in Zukunft eher geringere Priorität am MPI-M.

Michael: Würde Feedback der deutschen Community (aus DICAD) zu CMIP6 (im Hinblick auf CMIP7) in internationale Panels mitnehmen (WIP, ?).

Michael: Abschätzungen für CMIP6-Daten zu hoch (20 PB, momentan <10 PB) durch: Fehlen von Manpower, eingeschränkten Ressourcen (Rechenleistung, Speicherplatz), Unterschätzung des Arbeitsaufwandes (Standardisierung, ...)

Christopher, Michael, Christian: Ohne DICAD wäre CMIP6 deutlich schwieriger zu realisieren gewesen.

Christian: Möglichkeit, ähnliches Projekt für die Datenstandardisierung der CORDEX-CMIP6-Daten zu realisieren.
Michael: Absprache zwischen dir, Martin, Fabian und Martina. In CMIP5 gab es so ein Projekt, es gab zunächst Probleme bei der Definition eines Datenstandards/Datenrequests für CORDEX.
Christian: Dies sollte für CORDEX-CMIP6 zumindest schon recht weit fortgeschritten sein.

Anmerkung Martin: Danke sehr für eure Beiträge. Ich denke wir konnten recht gut ausmachen wo etwas schiefgelaufen ist und Verbesserungsbedarf besteht. Nicht nur in DICAD sondern auch allgemein in CMIP6. Für letzteren Punkt hat Michael angeboten unser Feedback in die versch. internationalen Panels zu tragen (WIP, …). Hierfür werde ich demnächst noch eine Mail rumschicken, sodass wir dieses Feedback vertiefen und ausformulieren können.

Allgemeine Diskussion:

- Erörterung ob eine Stoffsammlung (mit evtl. anschließender Diskussion) bzgl. der **technischen Organisation von CMIP6** gewünscht wird. Die daraus entstehenden Punkte könnte man an die Verantwortlichen weitergeben (bzw. über die WIP-Angehörigen der Partner), sodass dieses Feedback bei der Planung von CMIP7 evtl. berücksichtigt werden kann.
- **Aus CMIP6 Analysis Workshop, March 2019: How can we do better?**
    - Ensure **timely delivery** and enhanced quality control for the **forcings** through program-level support-More continuous, ongoing and more institutionalized (avoid single-point of failures)
    - The growing dependency on CMIP products by a broad research community and by national and international climate assessments means that basic CMIP activities, such as the **creation of forcing** datasets, the provision and archiving of CMIP products, and model development, require substantial efforts that **must be better funded**.
- **CMIP Continuity**:
    - Separation of the timescales for
        - "CORE EXPERIMENTS" REQUIRED FOR USERS (DECK, historical + possibly others)-Can go on faster timescales-More **automatic infrastructure** in place through program level support (e.g. for forcings), also at the modelling centers (to reduce the burden)
        - RESEARCH (CMIP6-Endorsed MIPs)-Infrastructure also needs to support the CMIP6 Research Activities-Could go on longer timescales
    - We have defined **enough experiments** and research questions in CMIP6 **to fuel research over the next phase** (CMIP7 fully building on CMIP6-Endorsed MIPs)
- **Allgemeine Zukunftsplanung im Projekt sowie evtl. Ausblick auf CMIP7.**
  (für Ergebnis siehe Präsentation im Redmine)

Interne Diskussion:

(für Ergebnis siehe Präsentation im Redmine)

- Abschlussbericht
- Verlängerung TP 1 (DKRZ) und 4 (FUB)
- Status Verbund 2