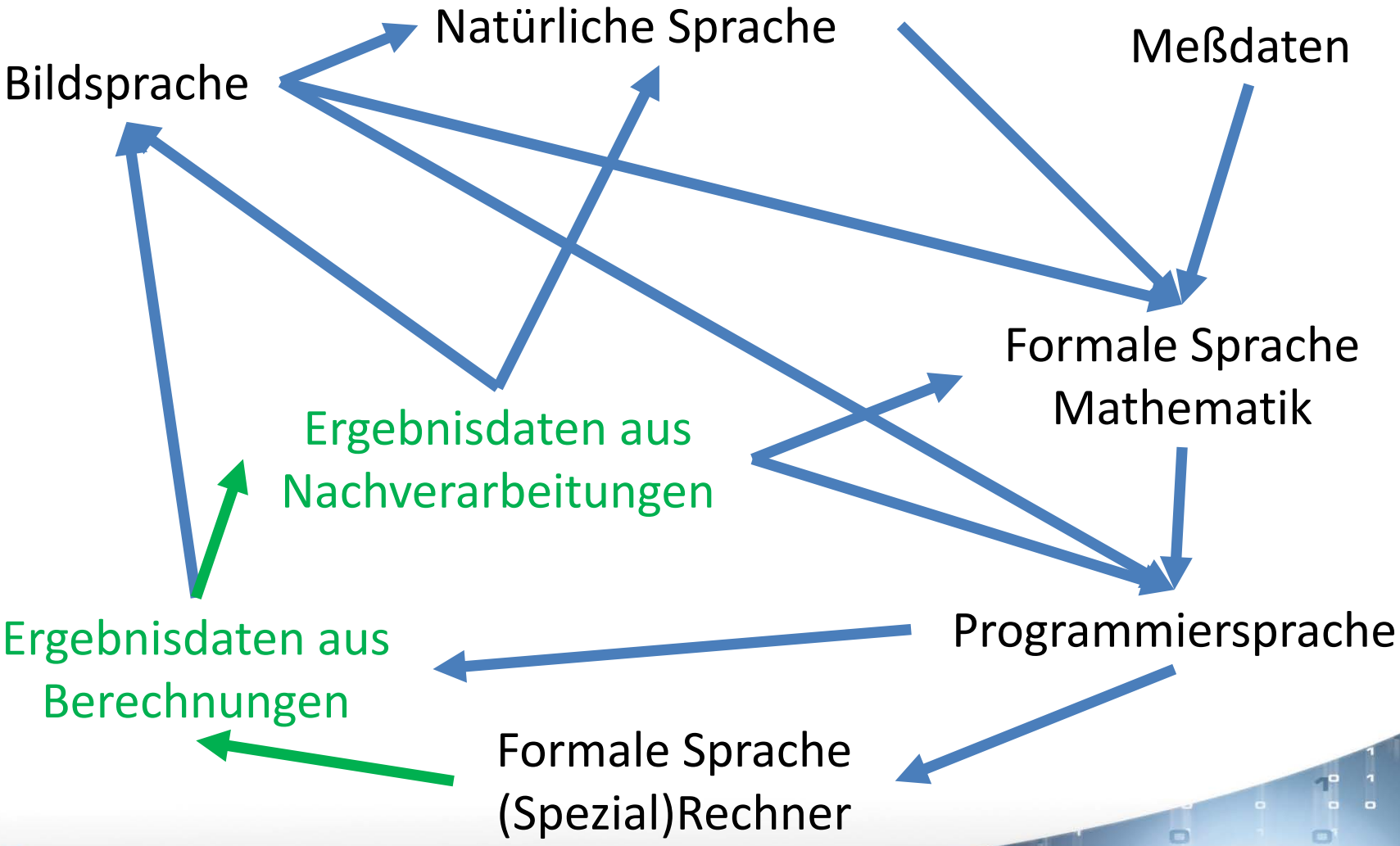
# Dialekte der Klimaforschung

## Vom parallelen Programm zu den Ergebnisdaten

Stephanie Legutke, DKRZ

# Prolog

## Vom parallelen Programm zu den Ergebnisdaten



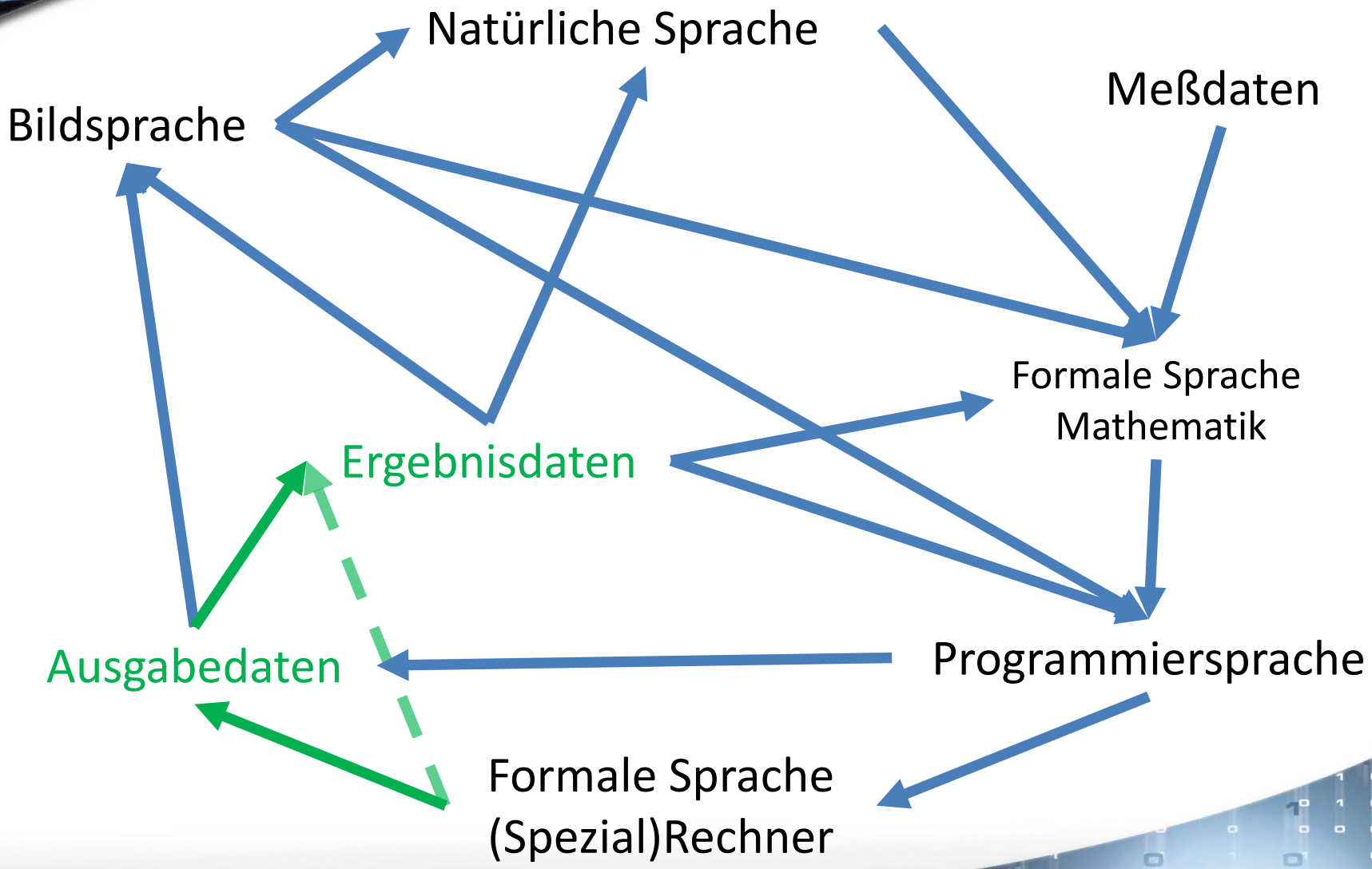
# Prolog

## Was sind „Ergebnisdaten aus  Berechnungen“?

| Ergebnisdaten aus den Berechnungen | Ausgabedaten (output data) | Rohdaten (model raw output) |
|---|---|---|
| haben Aussagepotential | FORTRAN WRITE | benötigen Verarbeitung |

***=> die Bezeichung ´´Ausgabedaten´´ macht keine Annahmen***

# Prolog


Natürliche Sprache
Bildsprache
Meßdaten
Formale Sprache
Mathematik
Ergebnisdaten
Ausgabedaten
Programmiersprache
Formale Sprache
(Spezial)Rechner

# Prolog

Vom parallelen Programm zu den Ausgabedaten

Von den Ausgabedaten zu den Ergebnisdaten

Die Abkürzung ‘Vom parallelen Programm zu Ergebnisdaten’

# Prolog

**Die Sprachebenen in der ‚Klimaforschung oder -modellierung' sind abhängig von den Skalen und Modellklassen, und den Ergebnisse, an denen man interessiert ist.**

**Beispiel: globale / regionale Klimamodelle**

| | Globale Klimamodelle | Regionale Klimamodelle |
|---|---|---|
| externer Antrieb | Solarstrahlung, Vulkanaerosole | Globalmodellparameter |
| Ausgabedaten | 6hr | 1hr |
| Ergebnisdaten | Niederschlagmonatsmittel | trockene Tage (< 1mm), Starkregen (> 50 mm/d) |
| Modellentwicklung | Kontrollläufe (keine Variabilität im externen Antrieb) | Evaluierungsläufe (Antrieb mit Reanalysedaten) |
| lange Läufe | ➢ O(1000 Jahre) | ➢ O(100 Jahre) |

# Prolog

**Alles Folgende**

- bezieht sich auf globale Klimamodelle und die Skalen, die damit typischerweise untersucht werden
- ist von der zur Zeit laufenden Erzeugung und Verarbeitung der Daten fürs CMIP5 (IPCC/AR5) geprägt, d.h. Beschreibung eines Ist-Zustands

Der Ist-Zustand passt sich laufend den Fortschritten der Rechen- und Speicherkapazitäten an.
Die Anpassung (an die R/S-Kapazitäten) kann durch höhere Auflösung, mehr/längere Läufe, oder auch durch eine Erhöhung der Modellkomplexität erfolgen (Beispiel: Klima- -> Erdsystemmodell)

# Vom parallelen Programm zu den Ausgabedaten

- Welche Ebenen erfasst die Transformation?
- Wozu diese Transformation?
- Was ist einfach?
- Was ist schwierig?
- Welche Varianten gibt es?
- Wie sieht die Praxis aus?
- Wie gehe ich vor?
- Was muss ich wissen?
- Wie kontrolliere ich die Korrektheit?
- Was bringt die Zukunft?

# Welche Ebenen erfasst der Übergang?

**Ausgangsebene:**

Modellrechnung, Simulation, diskretes paralleles Programm, (parallele) Ausgaberoutinen

**Zielebene:**

Ausgabedaten, die ‚den Anforderungen' genügen;

auf Platte, auf Band; tragbare externe Datenträger;

im schnellen Zugriff;

**Methode:**

FORTRAN WRITE bzw. entsprechende MPI/OpenMP Anweisungen; oder höhere Programmlayer (e.g. NetCDF, GRIB ); (paralleles I/O;)

# Zielebene **Ausgabedaten**

Rein technischer  Begriff.

Attribute:

- Parameterliste
- Frequenz (per Parameter)
- Dateiformat
  - z.B. NetCDF, komprimiert
- Datenformat
  - Big/Little Endian/IEEE
  - Genauigkeit
- Volumen
- Dateiname

# Zielebene **Ausgabedaten**

Macht a priori keine  inhaltliche Aussage,
und kann enthalten

- prognostische Variablen
- diagnostische Variablen
- instantane , akkumulierte Werte im Ausgabeintervall
- Extrema im Ausgabeintervall
- globale, zonale, vertikale Mittel von 3D Variablen
- komplexe Diagnostiken

# Wozu diese Transformation?

**Ziel:**
Gewinnung von **Ergebnisdaten**, die Erkenntnisse über das Klimasystem (oder ErdSystem) erlauben

**Notwendigkeit der Ebene:**
Vor dem Modelllauf steht oft nicht genau fest, welche Ergebnisdaten später nützlich sind. Deshalb wird (meistens) die Ebene **Ausgabedaten** zwischengeschaltet.
Bei Ausgabe von komplex diagnostizierten Ergebnisdaten verliert man eventuell Flexibilität für spätere Diagnostiken.

# Was ist einfach?

- Für größtmögliche Flexibilität, ist das Optimum
  - jeder Rechenpunkt (Zeit,Raum)
  - jede (prognostische) Variable

- Die Methode ist einfach:
  *FORTRAN WRITE (oder ähnliches)*
  Parallele Ausgabe: Thema 3

# Was ist schwierig?

**Limitierung:**

- I/O Performanz, Speicherkapazität

***=> Reduktion vor Ausgabe nötig.***

**Limitierung:**

- Analysepotential (Wissenschaft)
- Potential für Ergebnisdaten (Wissenschaft und Gesellschaft)
- Statt Aussagekraft: Potential für aussagekräftige Ergebnisdaten

***=> Welche Ausgabedaten sind nötig? (Mittelschwer)***

***Das Optimum der Aussagedaten mit diesen beiden Nebenbedingungen festzulegen, ist schwierig.***

# Strategien zur Ausgabereduktion: Welche Varianten gibt es?

- nicht jeder Gitterpunkt
  gröberes Model, zonal. Mittel, etc
- ✓ nicht jeder Zeitschritt
  Ausgabeinterval = N x dStep
- ✓ nicht jede Variable
- ✓ Diagnostik im Modell (s.u. Thema 4c)
  zonal. Mittel, vert. Integral, etc.
- kürzere Läufe
- weniger Läufe
- regional variierende Ausgabefrequenz / Parameter
- zeitlich variierende Ausgabefrequenz / Parameter

# Wie sieht die Praxis aus?

## Ausgabeintervall bei MPI-ESM

| Ausgabeinterval | Ausgabe-parameter (Beispiele) | Gewünschte Auflösung | Modell-Beispiel MPI-ESM-LR (CMIP5 Version) |
|---|---|---|---|
| 6 hr (N=36) | alle | Tagesgang | **ECHAM6 Atmosphäre dStep=10min** |
| Tageswerte (N=20) nicht adäquat !?! | Max. Deckschichttiefe | Max/Min/Mittel am Tag (Schwellenwertfkt) Impuls/Sporadisch: Tiefenwasserbildung | **MPIOM Ozean dStep = 1h12min** |
| Monatswerte (N=600)<br><br>Jahreswerte (N=7200) | Wind stress<br><br>Primary Carbon Production by Phytoplankton | Jahresgang<br><br>langjährige Tendenzen (Kohlenstoffzyklus) | **MPIOM Ozean dStep = 1h12min** |

# Wie sieht die Praxis aus?

## ‚Shapes' der **Ausgabedaten** bei MPI-ESM

| Ausgabeformat | Ausgabeparameter | Modell-Beispiel MPI-ESM-LR (CMIP5 Version) |
|---|---|---|
| Spektral (wenig Speicher)<br>Gitter 2D<br>Gitter 3D (integriert)<br>Gitter 3D | Temperatur<br><br>Niederschlag<br>Wolkenwasser<br>Luftfeuchte | **ECHAM6**<br>**Atmosphäre**<br>**(< 1.9°/205 km)**<br>**GRIB** |
| 0D<br><br>Gitter 2D<br>Gitter 3D (integriert)<br>Gitter 3D | Sea Ice Mass Transport Through Fram Strait<br>Global Mean Sea Water Salinity<br>Sea surface elevation<br>Ocean_heat_x_transport<br>Vertical mass transport | **MPIOM**<br>**Ozean**<br>**(< 1.5° )**<br>**NetCDF** |

# Wie gehe ich vor?

Parameterliste,-frequenz,cell_method der **Ausgabedaten** bestimmen sich aus den gewünschten Ergebnisdaten: (z.B.)

- 6 hr instantan für RCM Antrieb
- Tagesmax, -min für Impaktforscher
- Kohlenstoffreservoirs für ES Forschung

Je genauer die Ergebnisdaten festgelegt werden können, desto besser können die Ausgabedaten festgelegt werden.
(*Das ist u.U. schwieriger für wissenschaftliche Forschung, wenn sich erst nach dem Lauf der Bedarf an mehr/anderen Ergebnissen ergibt.)*

... unter Berücksichtigung
des Speicherplatz
des Performanzreduktion bei I/O

# Was muss ich wissen?

- welches Speichervolumen zur Verfügung steht
- welche Ergebnisdaten gefordert und/oder welche Fragestellungen relevant sind
- erforderliche raum- und zeitliche Auflösung
- welche diagnostischen Parameter schon im Modell berechnet werden müssen:
  sei <> cell_method (instant, mean, min, max, accu, …),
  p und v prognostische Parameter, d(p,v) diagnostischer Ausgabeparameter und d(<p>,<v>) \= <d(p,v)>
  => <d> Ausgabeparameter.

# Was muss ich wissen?

Beispiele diagnostischer Ausgabeparameter:

- Windgeschwindigkeit (speed): Austausch an Grenzflächen)
- Wärmetransporte: $v_{polwärts}(x,y)$ und T(x,y) sind korreliert
- Dichte: $\rho$ (T,S) nicht linear
- ‚vertical Adjustment'
- ‚clear sky' Diagnostik

# Wie kontrolliere ich die Korrektheit?

- Im allgemeinen ist es nicht schwierig, fehlerfrei (Gitter)Daten auszugeben.
- Korrektheit der Ausgabedaten entspricht i.A. der Korrektheit des Modells.
- Benutzung von Ausgaberoutinen vorgeben (Thema 3)
- Ergebnisdaten mit komplexer Diagnostik können eher falsch sein.

# Was bringt die Zukunft?

- Mehr Plattenplatz als RZ ?
- Cloud storage?
- Paralleles I/O?
- Effiziente Libraries um Projekt-Ergebnisdaten direkt auszugeben?
  - z.B. CMOR für CMIP5
  - z.B. CDIs fit machen

  Fehlbedienung möglich!

# Von den Ausgabedaten zu den Ergebnisdaten

# Von den Ausgabedaten zu den Ergebnisdaten

- Welche Ebenen erfasst die Transformation?
- Wozu diese Transformation?
- Was ist einfach?
- Was ist schwierig?
- Welche Varianten gibt es?
- Wie sieht die Praxis aus?
- Wie gehe ich vor?
- Was muss ich wissen?
- Wie kontrolliere ich die Korrektheit?
- Was bringt die Zukunft?

# Welche Ebenen erfasst der Übergang?

**Ausgangsebene:**

Ausgabedaten; eventuell schon als Ergebnis;

auf Platte, auf Band; im schnellen Zugriff;

**Zielebene:**

Daten, die direkt Erkenntnis über das Klimas ermöglichen (z.B. Klimasensitivität= 2.7 K );

entsprechend den Anforderungen;

auf Platte, auf Band; tragbare externe Datenträger; Datenbank (z.B. ESG)

im *praktischen* Zugriff für alle Teilnehmer (am Diskurs/Projekt);

**Methode:**

Offline (komplexe) Diagnostik;
Parameter nach Absprache (Betreuer, Arbeitsgruppe, Institute, Projekt)

# Wozu diese Transformation?

**Ziel sind Ergebnisdaten**, die

- Erkenntnisse über das Klimasystem (ES) bringen
- Vergleich mit Beobachtungen (Evaluierung) erlauben
- bei der Erstellung von Messstrategien helfen
- Vergleich von Modellen (MIP Projekte) ermöglichen
- zur Visualisierung geeignet sind
- Verständnis bei Nicht-Experten erzeugen
- zur Kommunikation mit  Politik, Gesellschaft geeignet sind
- einprägsam sind ($\Delta T$=2.7 K)

# Was ist einfach?

Diagnostik wird einfach(er), wenn zur Verfügung steht:

- viel Memory, Speicherplatz mit schnellem Zugriff
- entsprechende (getestete) tools
  - z.B. 'cdo dv2uv ...‘, 'cdo sp2gp …’
- vernünftige/verlässliche Vorgaben
- exakte Fragestellungen

Einfach ist:

- Dimensionsreduktion im Modellgitter (x,y,z,time)
- Einheitenanpassung

# Was ist schwierig?

Die Schwierigkeit bei der Festsetzung der Spezifikation von Ergebnisdaten

- steigt mit der Größe der Gruppe / Anzahl der Modelle
- fällt mit der Erfahrung
  - Wiederholung eines Projekts führt oft zur Festlegung und Akzeptanz von Standards (CMIP3 -> CMIP5)

**Beispiel CMIP5**

ECHAM6: keine Probleme (lange Erfahrung mit AMIP, CMIP1-3)
MPIOM: neue Diagnostiken rel. CMIP3
JSBACH: grosse Probleme (teilweise war nicht klar, wie die Parameter berechnet werden); modellspezifische Parameter
HAMOCC: wie JSBACH

# Was ist schwierig?

- Diagnostik auf nicht-regulären Gittern :
  z.B. ICON Gitter: wie sieht das zonale (O/W) Mittel aus?
- Diagnostik auf nicht-geographischen Gittern:
  z.B. tripolares Gitter:
  wie wird der meridionale Wärmetransport berechnet?
- Komplexe Diagnostik:
  z.B. ISCCP Satelitensimulator für CFMIP

# Was ist schwierig?

- die Werkzeuge hinken der Entwicklung (MPP) hinterher:
  - sie skalieren nicht oder schlecht
  - man muss eigene Parallelverarbeitung einbauen
    - Monate, Parameter, Komponenten/Ausgabedateien
- Vergleich mit Beobachtungen zur Modellevaluierung sind schwierig wg. der unterschiedlichen Skalen
- Vergleich von Modellen kann schwierig sein, wenn nicht dieselben Parameter berechnet werden:
  - new sea ice production
  - Parameter des Vegetationsmodells
  - Parameter der marinen BGC

## RZ für 1 Jahr Simulation/Datennachbereitung eines CMIP5 Experiments

MR: auf T63L40/GR15L40 Gittern für ECHAM6/MPIOM
LR: auf T63L95/TP04L40 " " " "

| 1 Jahr Simulation | 1 Jahr Datenverarbeitung | |
|---|---|---|
| 827 CPUh | 11 CPUh | RZ für MR |
| 155 min | 40 min | WCT für MR |
| 3.9 | | MR-Faktor$_{WCT}$ |
| 75.2 | | MR-Faktor$_{RZ}$ |
| 192 CPUh | 5.5 CPUh | RZ für LR |
| 90 min | 20 min | WCT für LR |
| 4.5 | | LR-Faktor$_{WCT}$ |
| 35.0 | | LR-Faktor$_{RZ}$ |

# Welche Varianten gibt es?

- Diagnostik im Modell
- Gemeinsame tools
  - CDOs, afterburner für ECHAM6 Daten
  - CMOR für CMIP5 Partner
- Standard  Verarbeitung  in der run-shell

# Wie sieht die Praxis aus?

**z.B. CMIP5 mit MPI-ESM:**

- ECHAM
  - die meisten CMIP5 Parameter sind seit langem Ausgabeparameter
  - afterburner, CDOs
  - sonst einfache Diagnostiken für CMIP5
- JSBACH
  - neu in CMIPx
- MPIOM
  - viel Diagnostik im Modell
    - wenn fehlerhaft, eventuell offline im Skript
    - sonst offline im FORTRAN Programm
    - sonst gar nicht
- HAMOCC
  - nur einfache Diagnostik; unnötige Ausgaben

# Wie gehe ich vor?

- Kollegen, Projektpartner  fragen
  - gibt es tools ?
  - welcher Rechner ist am besten geeigent?
- hotline fragen (z.B beratung@dkrz.de)
- googlen

# Was muss ich wissen?

- Projektspezifikationen
  z.B. CMIP5: http://cmip-pcmdi.llnl.gov/cmip5/docs/standard_output.xlsx
- wie häufig wird die Diagnostik voraussichtlich gemacht ?
  => eventuell Skript schreiben
- welcher Prozess ist automatisierbar
  - ist Datums-Reihenfolge zu beachten
  - kann ich später Teilbereiche nachprozessieren?
- was sind die Ausgabedaten des Modells?

# Was muss ich wissen?

**Projektspezifikationen von CMIP5**

- Anzahl Parameter angefordert
  MPI-ESM liefert ca. 60 % der angeforderten Daten
- Anzahl Parameter mit MPI-ESM abgeliefert:
  $94_{ECHAM6} + 212^{*}_{CFMIP} + 33_{JSBACH} + 73_{MPIOM} + 83_{HAMOCC}$
- Experimente/Jahre für die die einzelnen Parameter angefordert sind

*die CFMIP Diagnostiken sind nicht immer verlangt

# Nicht-Determinismus

- Verlässlichkeit (confidence)/Unsicherheit
  - Terminologie s. IPCC/AR4
  - in Klimaprognosen entstehen Unsicherheiten durch
    - Szenarien: => Spannweite RCP26 – RCP8.5
    - Modellformulierung: => Modellensemble betrachten
      - Bsp CORDEX diskutiert, nur pdf*s des Modellensembles zu veröffentlichen
    - Abhängigkeit von Anfangsbedingungen:
      - Realisationen: Experimentensemble (1-3!)

*probability density function

# Wie kontrolliere ich die Korrektheit?

- Wie sehen die Parameter bei Anderen, in früheren Rechnungen, in Beobachtungen aus?
- Sind Abweichungen davon zu erklären?
- Schwarmintelligenz
  - CMIP5 Daten im ESG, das weltweit Zugriff erlaubt, und wo die Daten aller CMIP5 Teilnehmer gespeichert sind, werden von den CMIP5 Teilnehmern ‚qualitätsgeprüft', indem sie Daten vieler Modelle gemeinsam verarbeiten/visualisieren
  - nur möglich, wenn ... Suchen, Finden, Verarbeiten gleicher Daten aller Modell einfach möglich ist

# Was bringt die Zukunft?

- Mehr Plattenplatz als RZ?
- Schneller Zugriff auf Bänder?
- Cloud storage?
- Effiziente Tools um Projekt-Ergebnisdaten auszugeben?
  - z.B. wie CMOR für CMIP5
  - z.B. CDOs fit machen
- Paralleles I/O und tools, die auf den cpu-Domains arbeiten ?
- GRID Computing?

# Die Abkürzung ‘Vom parallelen Programm zu den Ergebnisdaten’

# Diagnostik im Modell

**Cons**

- Performanz
  insbesondere bei (mpp) skalare Rechnerarchitekturen
- Eine falsche Diagnostik kann u.U. nicht korrigiert werden;
- Reduzierte Flexibilität
- Gitterabhängige Diagnostik
  (Transporte durch Passagen im Ozean)

# Diagnostik im Modell

**Pros:**

- die Diagnostik wird ohnehin im Programm berechnet (e.g. Dichte)
- sie ist schwierig zu berechnen; über Programmcode haben die Programmierer die Kontrolle
- Datenausgabe wird reduziert
    - allerdings ist Ausgabe als 2D UND als 3D kontraproduktiv

# Suchen, Finden, Verarbeiten

CMIP5 Datenarchive und Metadaten:

| | | |
|---|---|---|
| WDCC/DKRZ | [Cera] | (IPCC DDC) |
| ESG | [DRS,CIM] | (CMIP5 Federated Archive) |
| | | Datenknoten: DKRZ, BADC, PCMDI, … |

- Institutsnamen (e.g. MPI-M)
- Modellnamen (e.g. MPI-ESM-[MR,LR,P])
- Experimentnamen (e.g. historical, rcp45, decadalYYYY)
- Frequenzen (e.g. yr,mon[Clim],day,6hr,3hr,subhr)
- Variablen-/Dateinamen (e.g. tas_...nc)
- NetCDF/CF

CIM (CommonInformationModel) Metafor EU FP7

# Gegenüberstellung: standard_name <-> long name

| standard_name | long_name |
|---|---|
| Controlled Vocabulary (CV) | frei wählbar; in CMIP5 zu CV gemacht |
| formaler Name; Suchbegriff in Datenarchiven | Suchbegriff im CMIP5 Datenarchiv |
| weltweit in der Klimaforschungsgemeinschaft akzeptiert | in CMIP5 akzeptiert |
| benutzt in Datenvergleichen | benutzt z.B. in Graphik Überschriften; kann für Zielgruppe angepasst werden |
| build-rule: surface_..., ,..._where_sea_ice, etc. | |
| am PCMDI definiert | auf Projekt-, Institutsebene definiert |
| träger Prozess, da Zustimmung ‚weltweit' nötig | Zustimmung nur ‚projektweit' nötig |

# CMIP5 Beispiele für standard_name & long_name

| standard_name | long_name |
|---|---|
| sea_ice_transport_across_line | Sea Ice Mass Transport Through Fram Strait |
| surface_snow_thickness | Snow Depth |
| sea_water_salinity | Sea Water Salinity |
| tendency_of_mole_concentration_of_particulate_organic_matter_expressed_as_carbon_in_sea_water_due_to_net_primary_production | Primary Carbon Production by Phytoplankton |
| ocean_heat_x_transport | Ocean Heat X Transport |
| water_evaporation_flux_from_canopy | Evaporation from Canopy |
| area_fraction | Fraction Fraction of Grid Cell that is Land but Neither Vegetation-Covered |
| area_fraction | Total Primary Evergreen Tree Cover Fraction |

**long_name** natürliche Sprache (Modell-/Projektspezifisch)

**standard_name** Formale Sprache/Naturwissenschaften (wissenschaftsübergreifend akzeptiert in der