Met Office
Hadley Centre

# Multi-platform data delivery workflows for CMIP6 and beyond

Matthew Mizielinski, Emma Hogan,

Piotr Florek, Mark Elkington

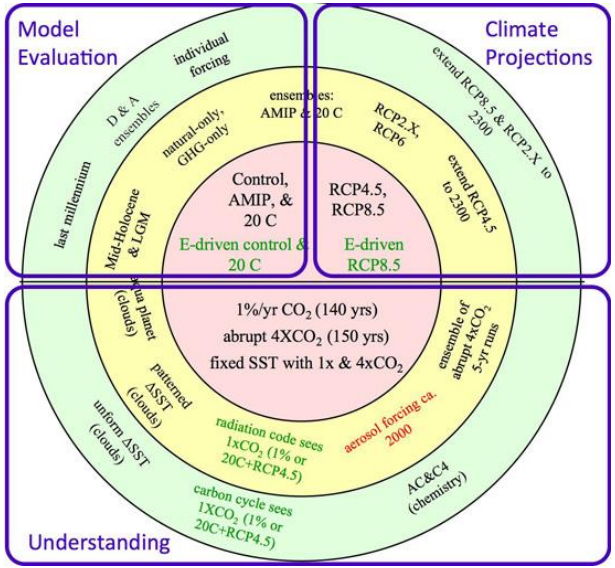
Met Office
Hadley Centre

# Outline

- Background to CMIP5 & 6
- The task for CMIP6
- Climate Data Dissemination System (CDDS) structure
- CDDS outside the Met Office: JASMIN
- CDDS Beyond CMIP6

Met Office Hadley Centre

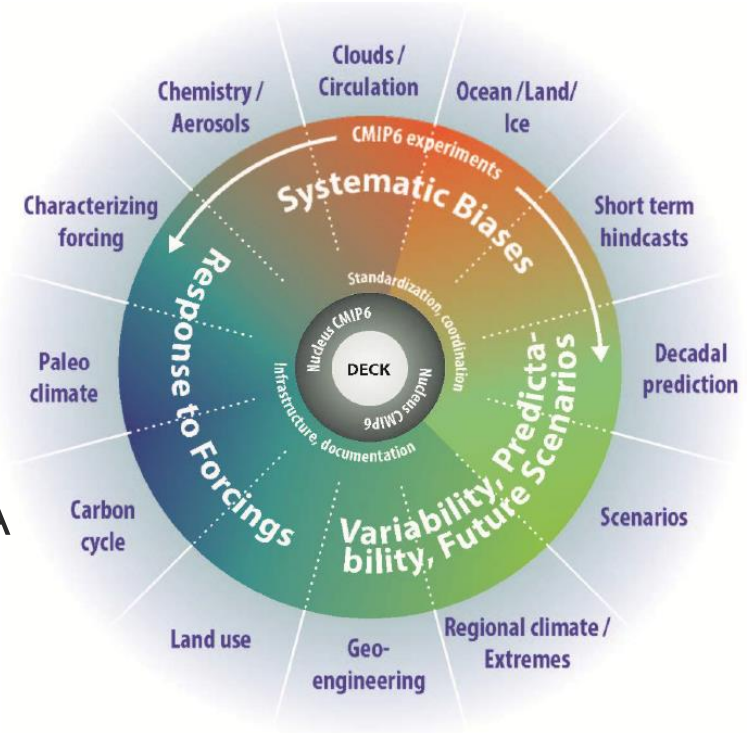
# CMIP5 background




- Around 30 experiments
- Met Office submission of 80 TB output data for 4 models (3 HadGEM2 variants plus HadCM3) across ~180 simulations
- Data produced on MO HPCs and archived to tape
- CREM/DDS (Mark Elkington & Jamie Kettleborough) ran on two dedicated servers each processing ~60GB/day
- Data copied to British Atmospheric Data Centre for archiving and ESGF publication
- CREM/DDS used for CORDEX, CCMI, QBoI, CLIVAR HWG

Met Office
Hadley Centre

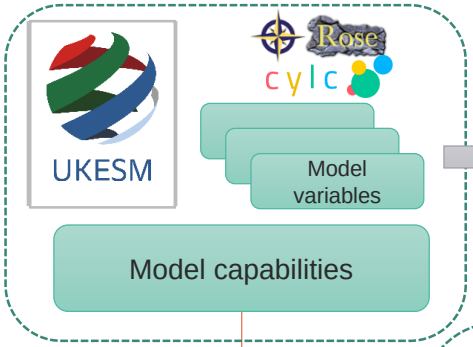
# CMIP6




- 20+ MIPs owning 250+ experiments
- Wider range of user communities
- 3 core models: 2 x HadGEM3 + UKESM1
- Models used by MOHC, NERC, NIWA, KMA
- Core submission expected to be ~1.5 PB (other MIPs extra)
- Need to be able to process data
  - On multiple platforms; MO and JASMIN for academic and international partners
  - In a reasonable time frame
- Some requirements clearly specified, others less so

# Climate Data Dissemination System (CDDS)

Met Office Hadley Centre

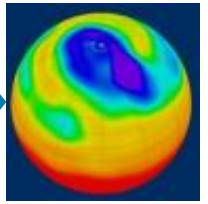



PP

netCDF

Archive (MASS)

Publication process



CEDA ESGF node

User Queries

CDDS

Suite information and metadata

Climate Research Experiment Management (CREM) database

ES-DOC Earth System Documentation

CMIP6 Data Request

CDDS-Prepare

Requested variables lists

CDDS-Extract

CDDS-Configure

CDDS-Convert

CDDS-QC

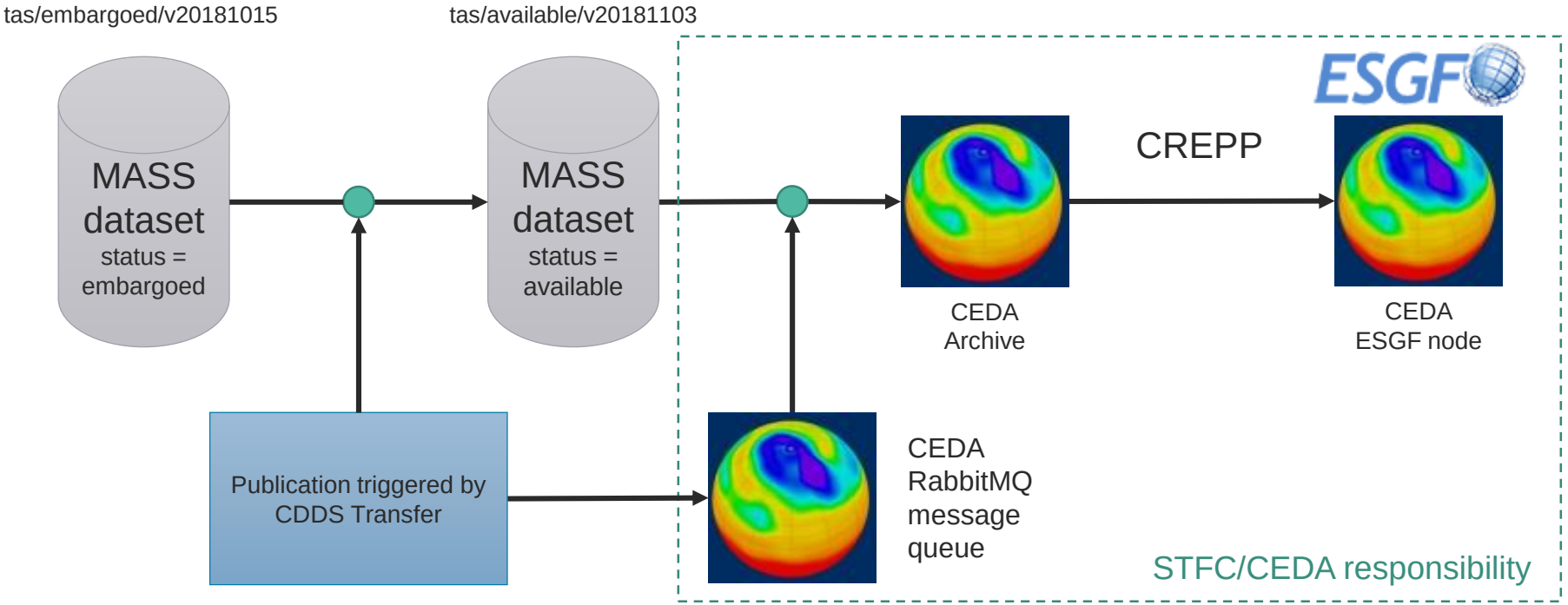
CDDS-Transfer

SPICE processing cluster

User / MIP requirements

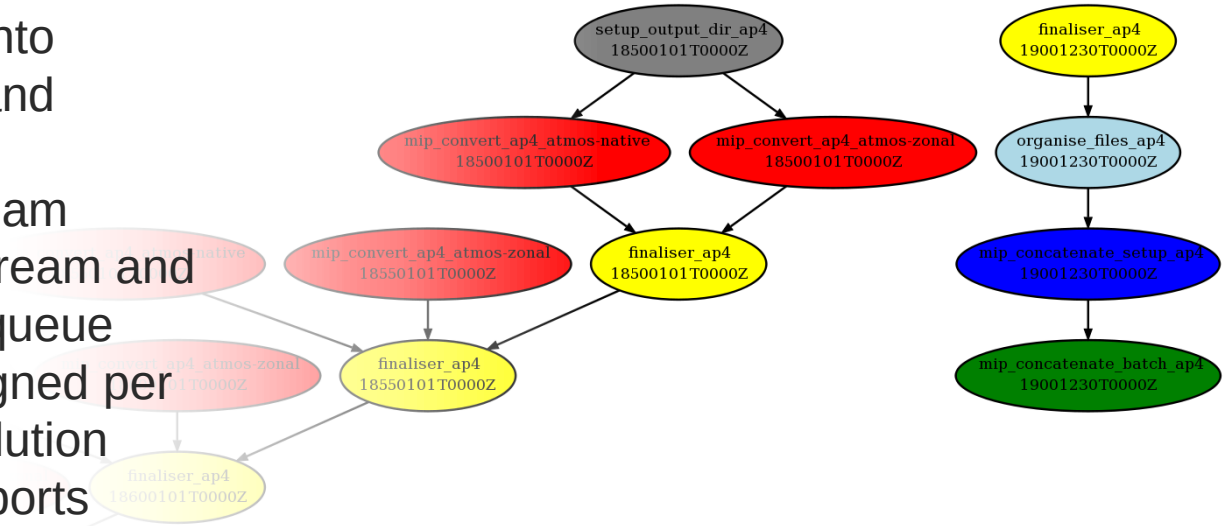
Met Office Hadley Centre

# Publication notification process



# Conversion suite structure

- Model data organised into streams by frequency and component
- Setup directory per stream
- MIP convert task per stream and “grid” limited by a cylc queue
- Cycling frequency assigned per stream and model resolution
- Finaliser per stream reports progress and triggers concatenation process
- `ncrcat` based concatenation




Cylc graph extract

Met Office
Hadley Centre

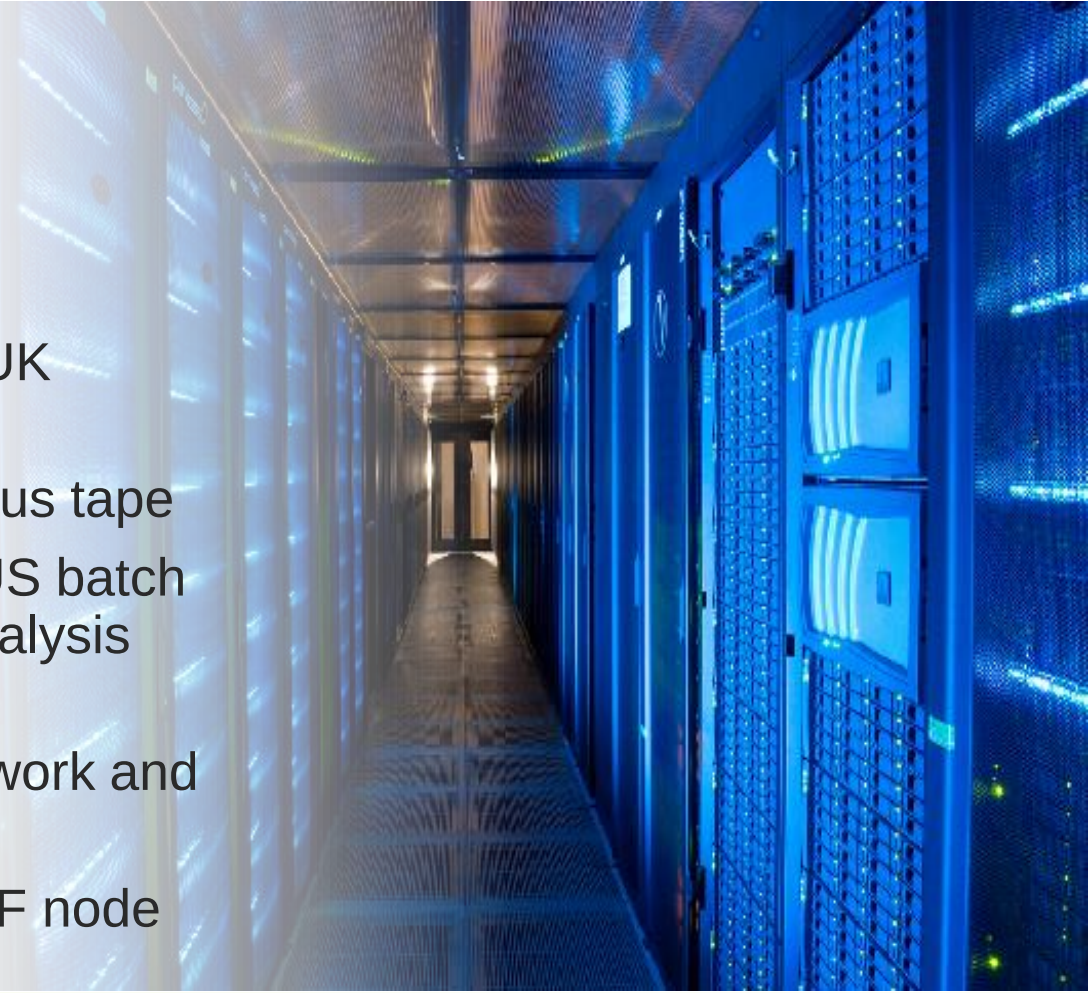
# The multi-platform bit

- We (MO) cannot do all data processing for everyone using HadGEM3/UKESM1 models
- NERC, NIWA and KMA to use JASMIN
- Not enough resources to support our database & admin interface tools for everyone
- Limited set of CDDS components planned to be used
- We (MO) do not want to be responsible for answering questions about data we did not produce

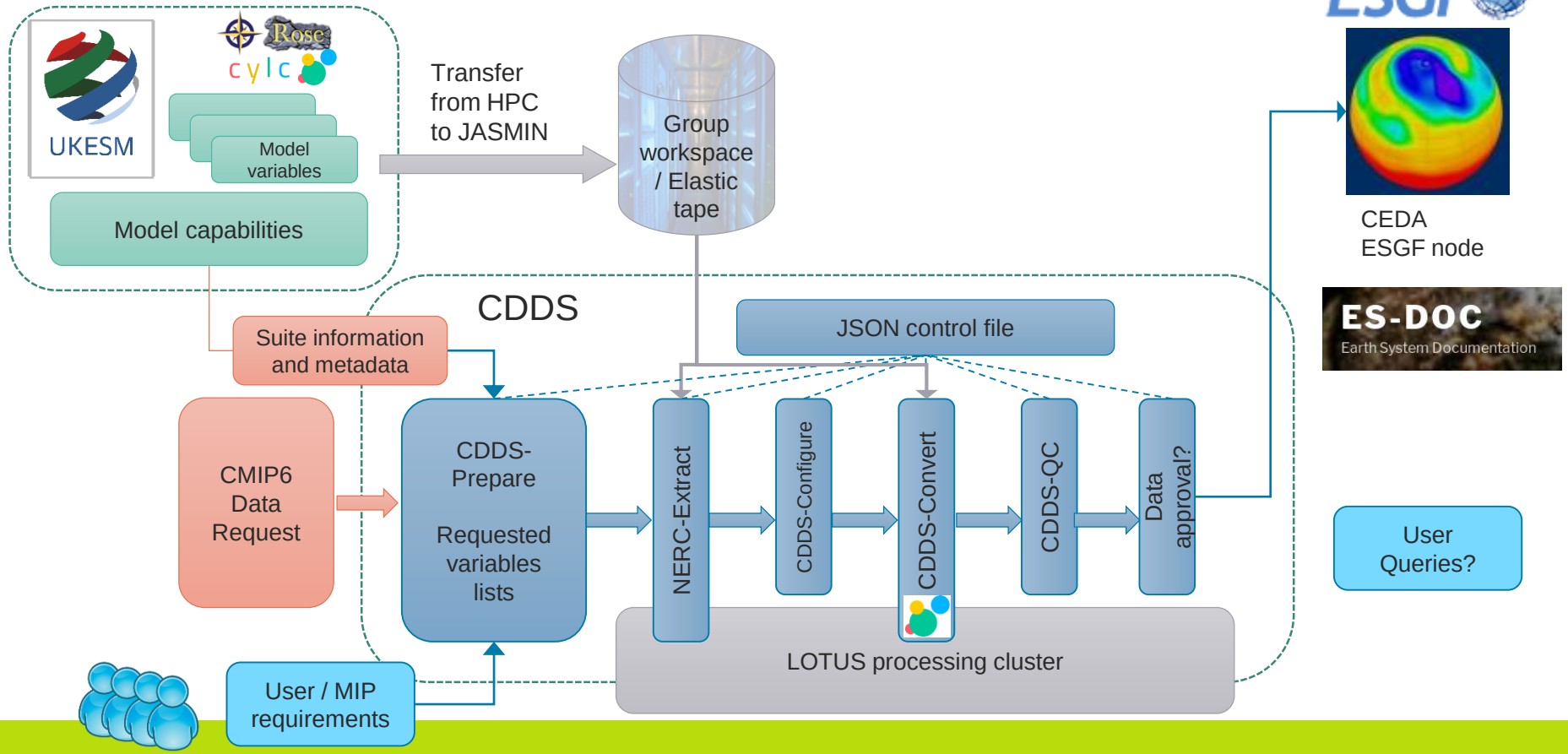
Met Office
Hadley Centre

# JASMIN

- Managed by STFC CEDA for UK environmental science
- 38 petabytes of disk storage plus tape
- 5,000 compute cores on LOTUS batch cluster plus interactive data analysis servers
- High-performance internal network and connections to the internet
- Hosts CEDA Archive and ESGF node



# CDDS Flow (JASMIN)



Met Office
Hadley Centre

# Phased delivery of data

Stage 1:  Monthly data (~400TB)

- All reviewed mappings available
- Any experiments ready

Stage 2: Daily data (~400TB)

- Introduce science users with some training
- One MIP at a time
- Catch up anything left over from stage 1

Stage 3: Sub-daily data (~800 TB)

- Everything else
- Ensemble class experiments (possibly stage 2)

Met Office Hadley Centre

Rose cylc

# Beyond CMIP6

- Greater Rose / Cylc integration for archive extraction, Quality Control and archiving
  - More efficient use of processing and storage resources
- Process routine model development simulations for analysis
  - Inefficient to maintain two ways of working; one for model development the other for inter-institute/international projects
- Configure models to give only the required output (stable data request)
- Grids, time slices and new models?
- Model output configuration through XIOS?

Met Office
Hadley Centre

# Summary

- Evolution not revolution for CMIP6
  - CDDS has a similar structure to CMIP5 system
- CDDS uses a Rose/cylc suite for the management of the conversion process
- Beyond CMIP6 requires more of a revolution;
  - Work towards a Rose/cylc suite for the whole of CDDS
  - Replace management database with an atomic management function
  - Support model development as well as data publication
  - New models (LFRic) and output configurations (XIOS)