Barcelona, 27/01/2016

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Big Data en Ciencias de la Tierra

Aplicación sobre una simulación climática en alta resolución

Workshop Red Supercomputación y eCiencia

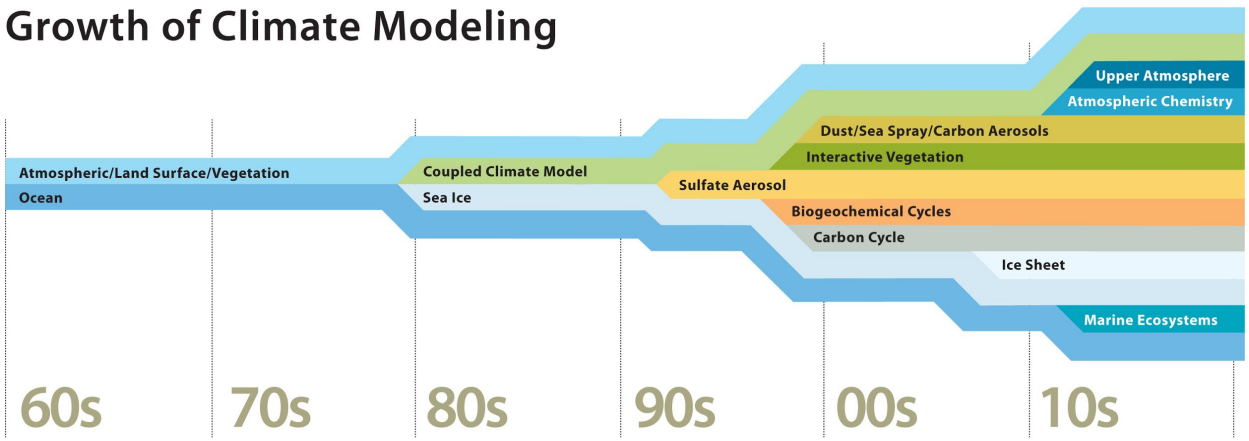
Pierre-Antoine Bretonnière
Earth Sciences Department

Plan

Los datos en una simulación climática en alta resolución:

1. El caso científico
2. El ciclo de vida de los datos
3. ¿Qué datos y qué volumen generamos?
4. Post-procesamiento: de formateo a visualización
5. Problemas encontrados debidos al aumento del volumen de datos y perspectivas

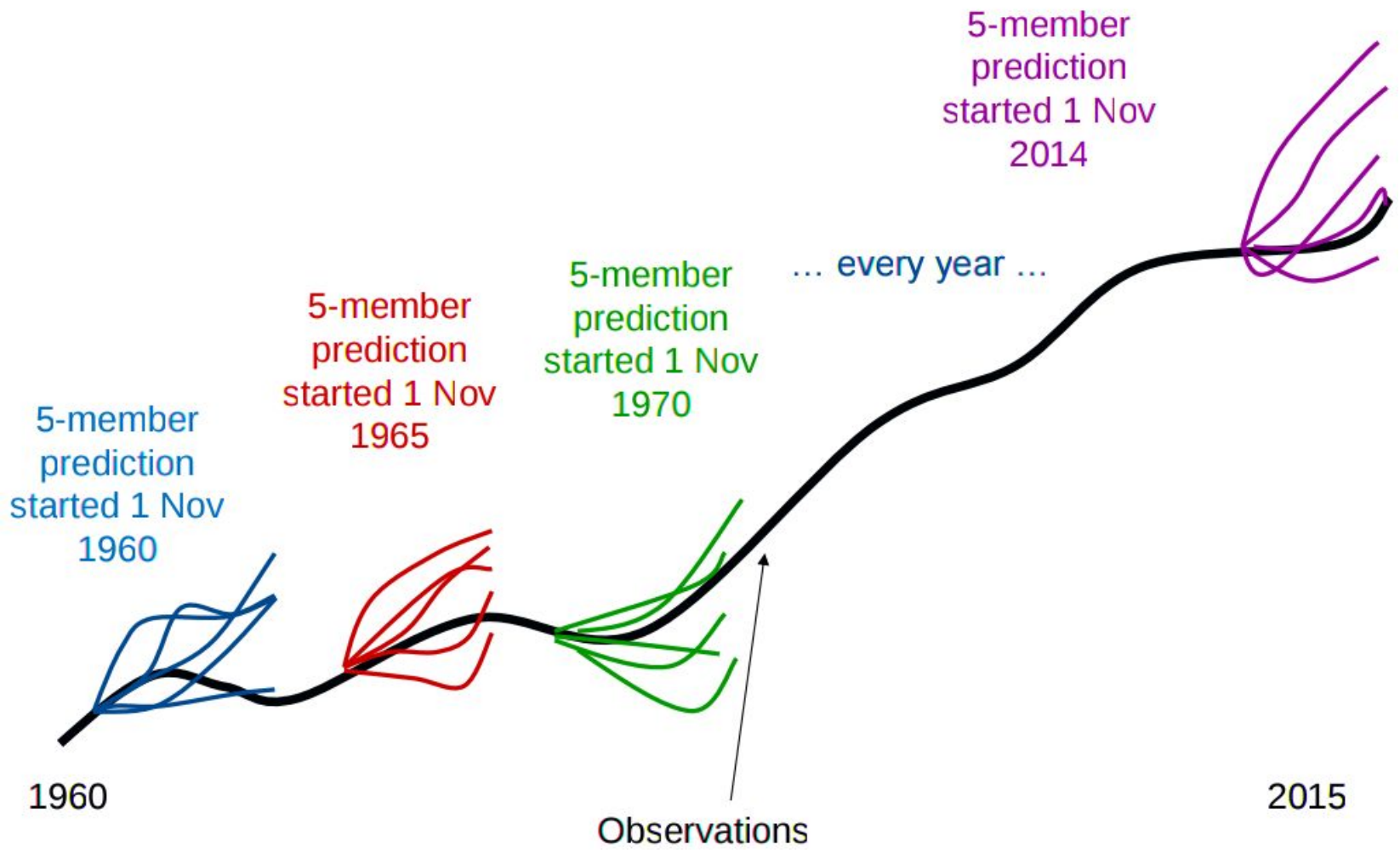
El caso científico



- Objetivo: predecir el clima en los próximos años (10 años -> siglo) mediante la comparación de los distintos GCM que han sido desarrollados por distintas instituciones a nivel global.

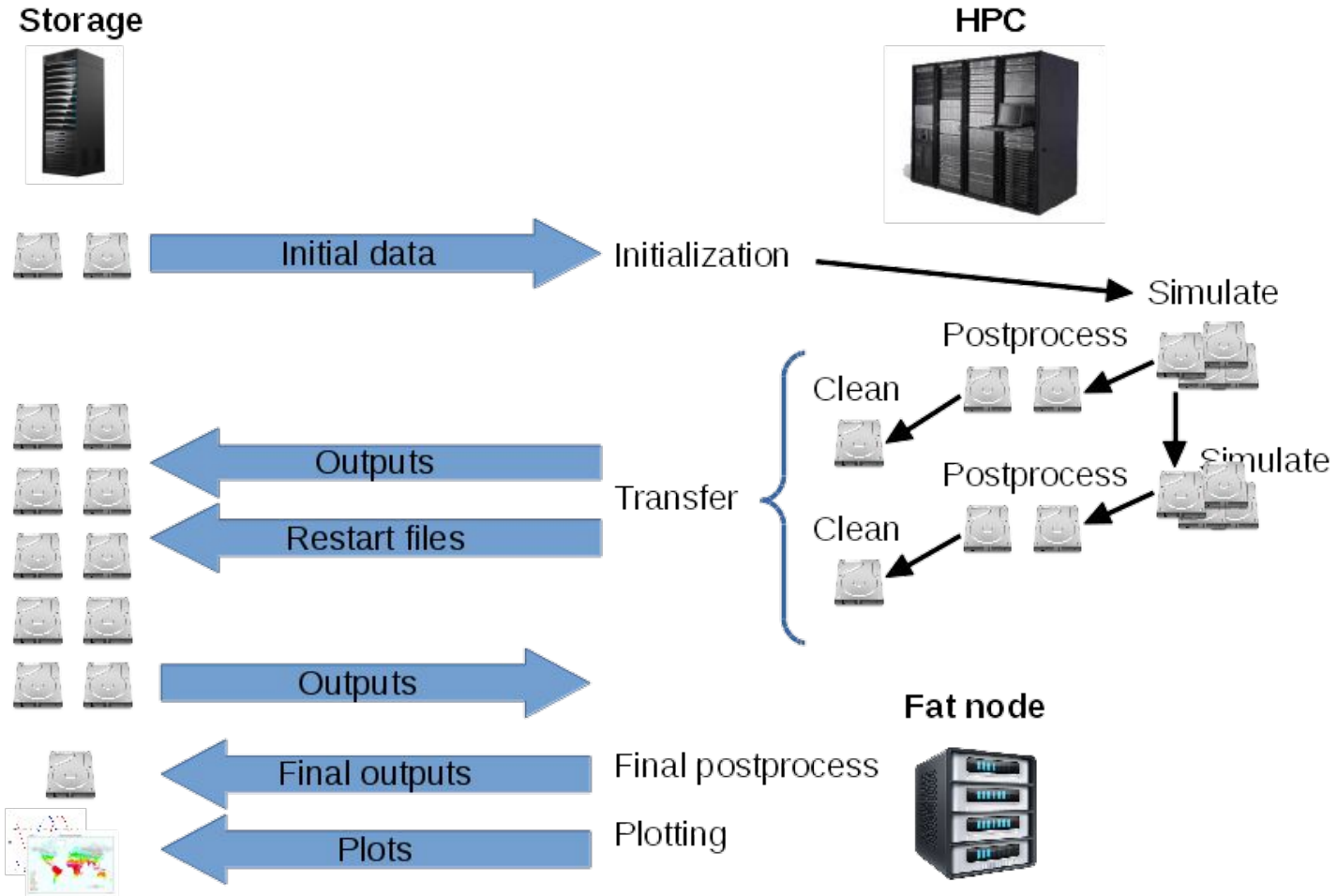
El caso científico

- Necesidad de inicializar los modelos a varias “start dates” y con distintas condiciones físicas iniciales



El ciclo de vida de los datos

- Necesidad de usar capacidades de HPC para correr los modelos pero también todo un workflow para generar y post-procesar todos los datos



Una idea de los datos generados

- Salidas de los modelos en ficheros NetCDF(3/4) o Grib(1/2) 3 o 4D
- Una simulación en alta resolución puede gastar 100.000 horas/año simulado/miembro

	Horizontal resolution (atmosphere/ocean)	**Output sizes of one year monthly simulation**
Standard Resolution	T255/ORCA1 60km/100km	26 GB
High Resolution	T511/ORCA025 40km/25km	120 GB
Ultra High Resolution	T1279/ORCA012 25km/12km	1TB

	CMIP	**CMIP2**	**CMIP3**	**CMIP5**
Number of experiments	1	2	12	110
Centers participating	16	18	15	24
Number of models	19	24	21	45
Total dataset size	1GB	540GB	36TB	3.3PB

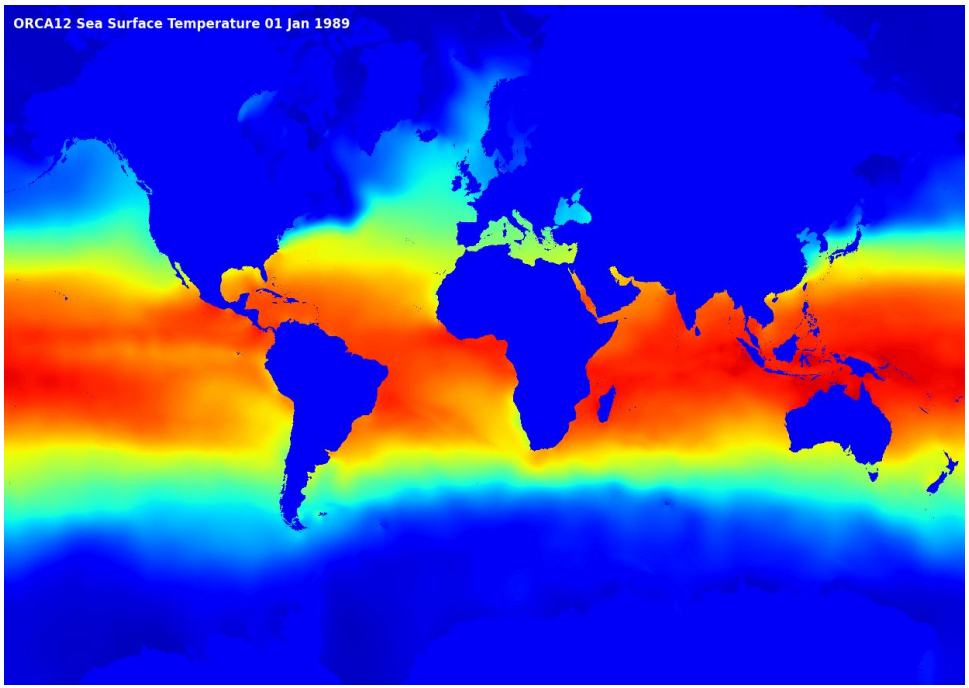
Post procesamiento de datos

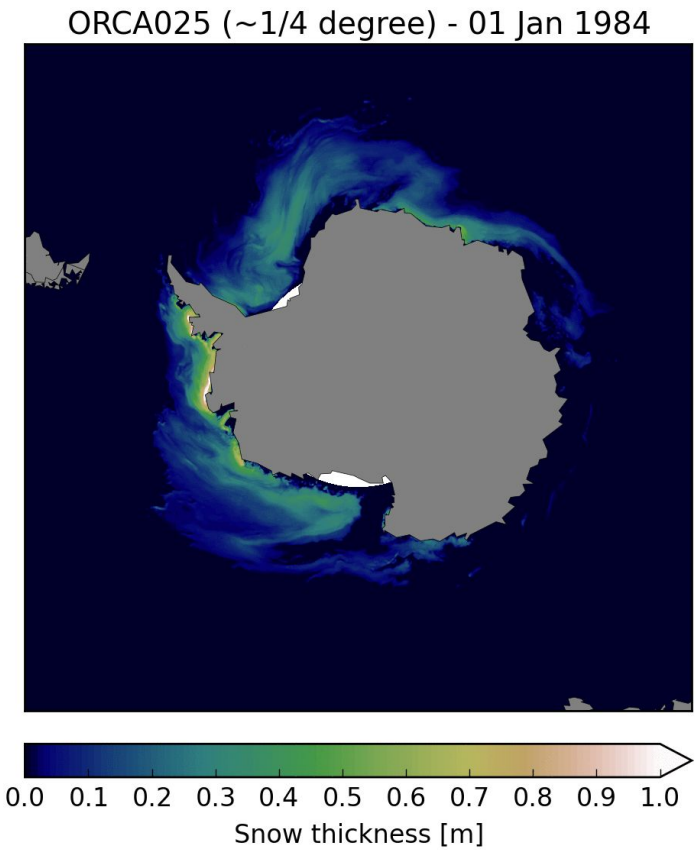
- El post-processing incluye tanto el **formateo** de los datos según ciertas convenciones establecidas (CMORización), como el **almacenaje** de los mismos en portales web para compartir los datos (ESGF) que han sido diseñados para compartir estos datos. Asociado a ello, surgen problemas de data discovery, subsetting y data indexing.

- El cálculo de variables derivadas (tasmin, tasmax, moc,...) y medias mensuales o anuales (climatologías) también requieren potencia de cálculo y uso intensivo de memoria.

Post procesamiento de datos

- La cadena de post-processing también incluye producción de visualizaciones.





Problemas identificados (y soluciones?)

- En toda esta cadena de operaciones típicas de una simulación climática, el aumento del tamaño y de las fuentes de datos implica, además de necesitar mucho más tiempo (processing > simulación), requiere **nuevas soluciones tecnológicas** para poder mejorar la investigación científica.

- A nivel técnico, tienen que adoptarse soluciones Big Data tanto para las herramientas (**Hadoop, Spark,...**) como para las metodologías (**“bring the compute to the data”**).

www.bsc.es

¡Gracias!

pierre-antoine.bretonniere@bsc.es