Progress in NWP on Intel HPC architecture at Australian Bureau of Meteorology

www.cawcr.gov.au

Robin Bowen Senior ITO

Earth System Modelling Programme

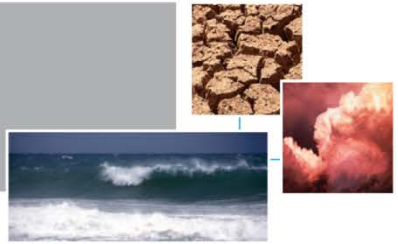
04 October 2012 ECMWF HPC

Australian Government
Bureau of Meteorology

The Centre for Australian Weather and Climate Research
A partnership between CSIRO and the Bureau of Meteorology

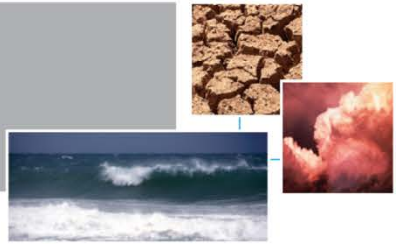
CSIRO

Presentation outline



- Weather and Climate research in Australia – BoM, CAWCR and ACCESS
- Current capability in NWP and Climate
- Recent NWP advances for severe weather
- HPC upgrades in next few years
- Benchmarking Sandy Bridge chips

CAWCR



Centre for Australian Weather and Climate Research

- Partnership between
  - Bureau of Meteorology
  - CSIRO Marine and Atmospheric Research (CMAR)

- Research and Operational Development focus on
  - Weather, ocean and seasonal prediction
  - Climate variability, change and impacts
  - Hazard prediction and warnings
  - Water supply and management
  - Responses to weather and climate related impacts

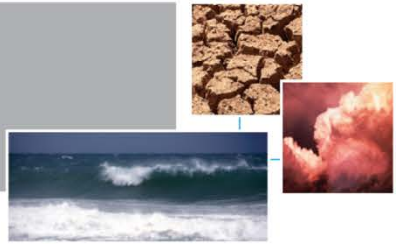
- Close links with
  - Weather and climate research groups across Australia
  - and across the world, for example Met Office, ECMWF

ACCESS

Australian Community Climate and Earth System Simulator

- Fully coupled system
  - Provide a *national* approach to climate and weather prediction model development
- Joint initiative
  - Bureau of Meteorology
  - CSIRO
  - Australian universities > access to common system
  - DCCEE - Australian Government Department of Climate Change and Energy Efficiency
- Focus on the needs of a wide range of stakeholders:
  - Providing the best possible services
  - Analysing climate impacts and adaptation
  - Linkages with relevant University research
  - Meeting policy needs in natural resource management

ACCESS Schematic



OBS

Assimilation (4DVAR)

Atmosphere (UM)

Dynamic Veg. (LPJ)

Assimilation

Land surface (CABLE)

Coupler (OASIS)

Chemistry (UKCA)

Ocean Carbon (CSIRO)

Ocean AusCOM (MOM4)

Sea-ice (CICE4)

Assimilation (BODAS)

OBS

NWP

Seasonal Climate

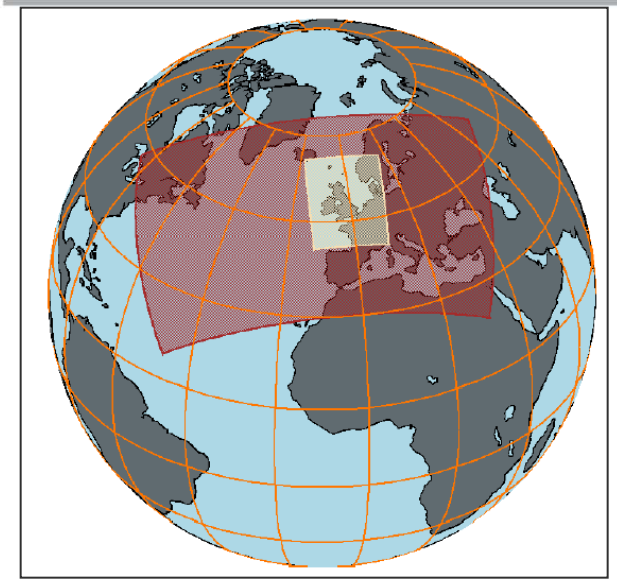
Climate Earth system

Australian Government
Bureau of Meteorology

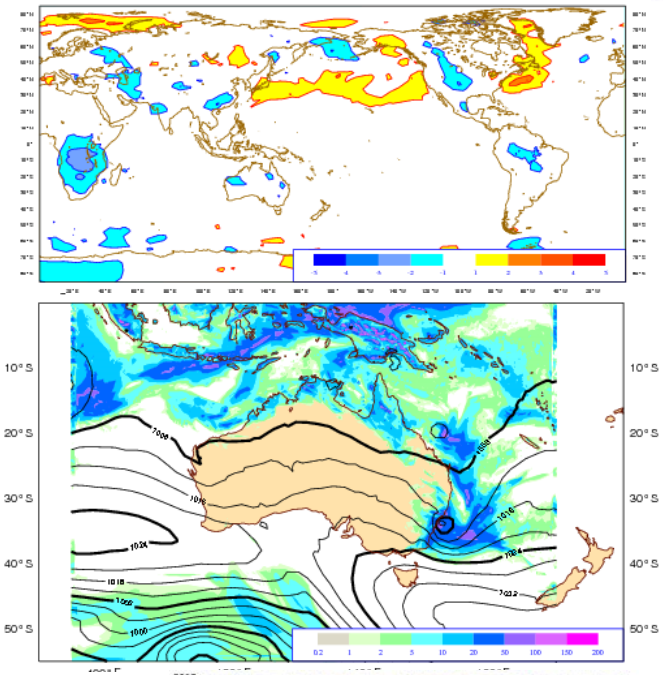
The Centre for A… e Research
A partnership be… Meteorology

CSIRO

ACCESS (BoM) NWP Domains



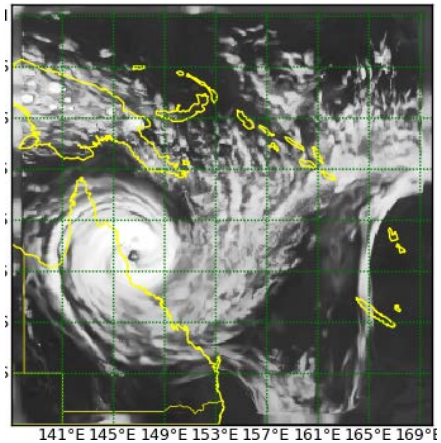
Unified Model



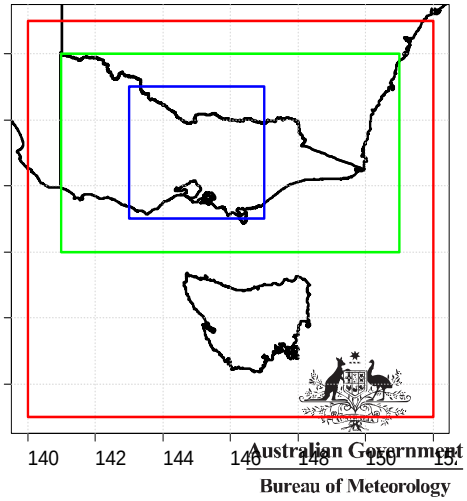
ACCESS-G
40km,
70 levels,
UM 7.5

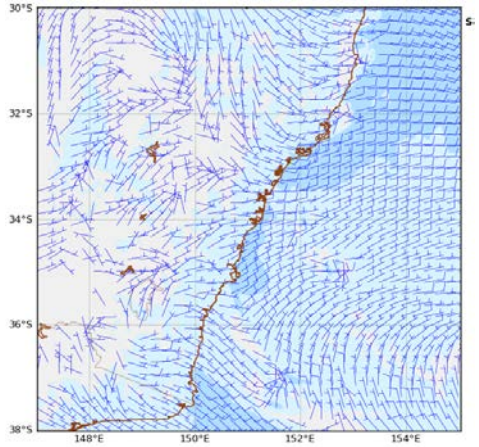
ACCESS-R
12km,
70 levels,
UM 7.5

ACCESS-TC 12km



ACCESS 1.5km
SREP, Fire Wx



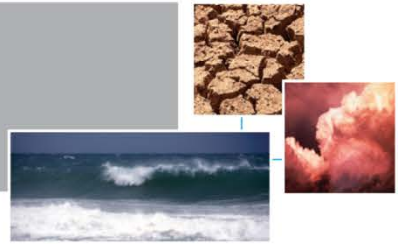


ACCESS-C
4km,
70 levels,
UM 7.5

The Centre for Australian Weather and Climate Research
A partnership between CSIRO and the Bureau of Meteorology

CSIRO

ACCESS (BoM) Global and Regional Ensemble Prediction System, AGREPS



AGREPS is being run routinely in research mode for Global and Australian Regional domains (from Met Office MOGREPS), since start of 2011

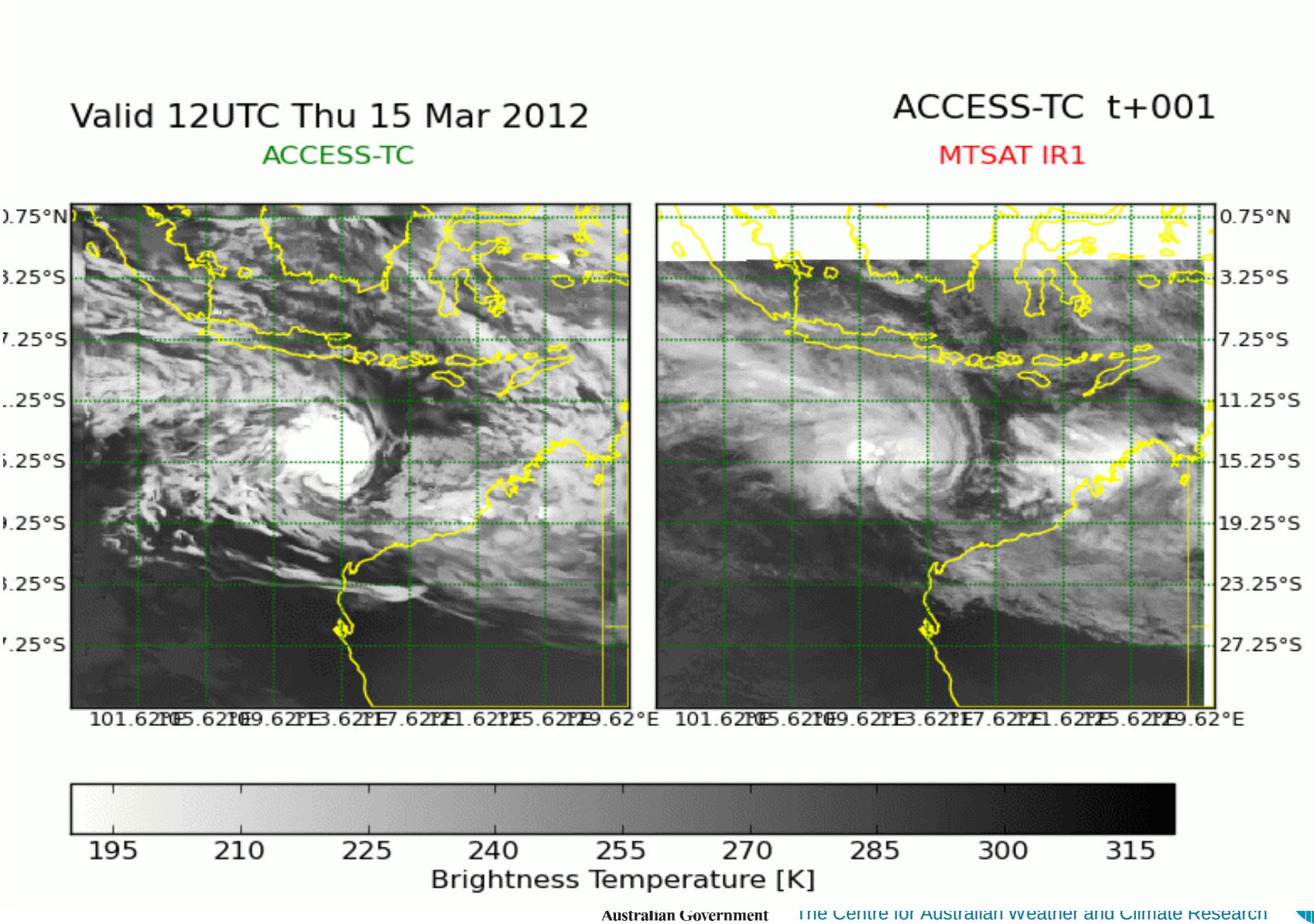
– APS0 AGREPS-G, N144 L50, 5day forc, using UM 6.4
AGREPS-R, 0.375 L50, 3 day forc, using UM 6.4

– APS1 AGREPS-G, N216 L70, 5day forc, using UM 7.9
planned AGREPS-R, 0.22deg L70, 3 day forc

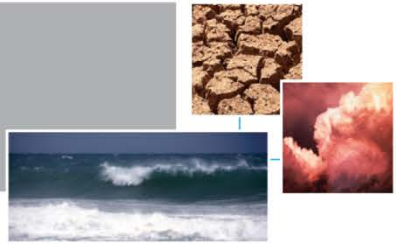
– 24 members, 23 perturbed + 1 control. Regional ensembles receive lateral boundary conditions from global ensemble

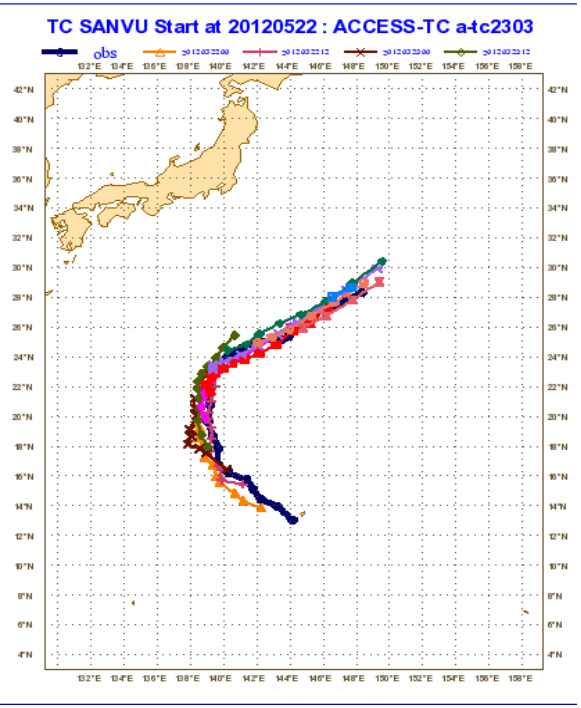
– Detailed evaluation is being performed. Resources for Operational implementation are not yet available.

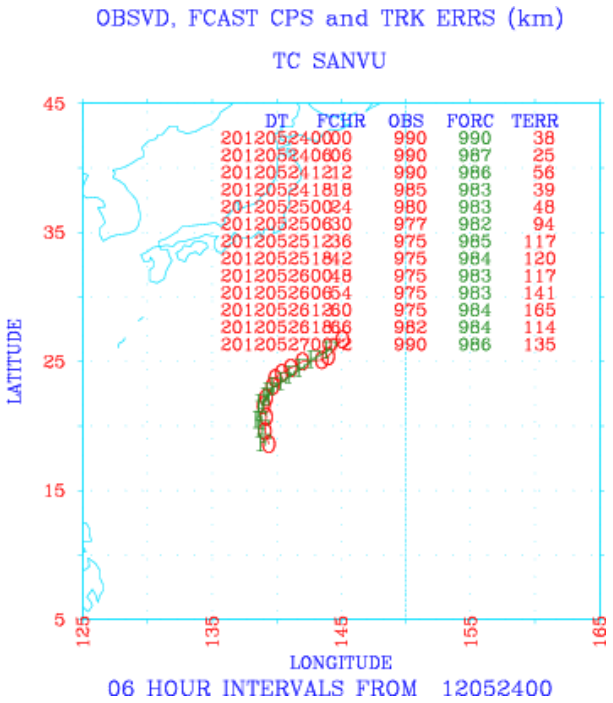
TC Lua
Base Date: 20120315 Base Time: 12UTC

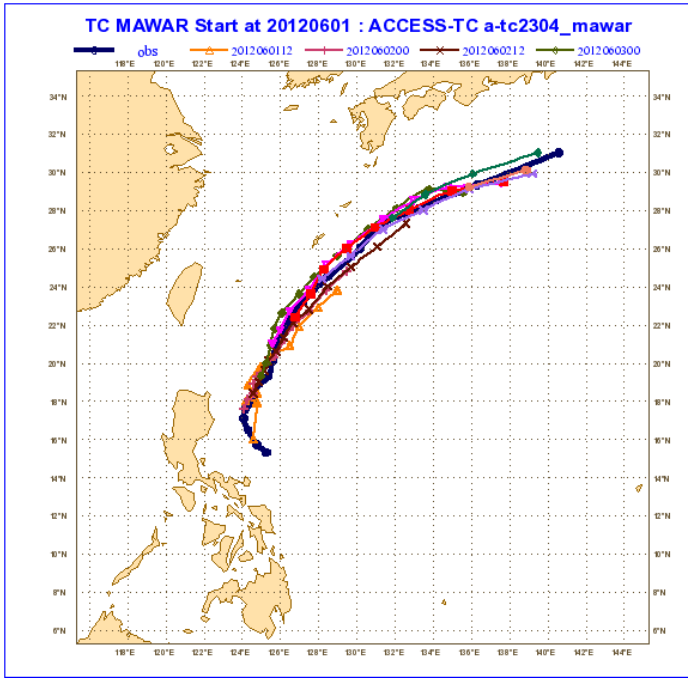


Operational ACCESS-TC Forecasts for SANVU (May 2012) and MAWAR (June 2012)









Australian Government
Bureau of Meteorology

The Centre for Australian Weather and Climate Research
A partnership between CSIRO and the Bureau of Meteorology

CSIRO

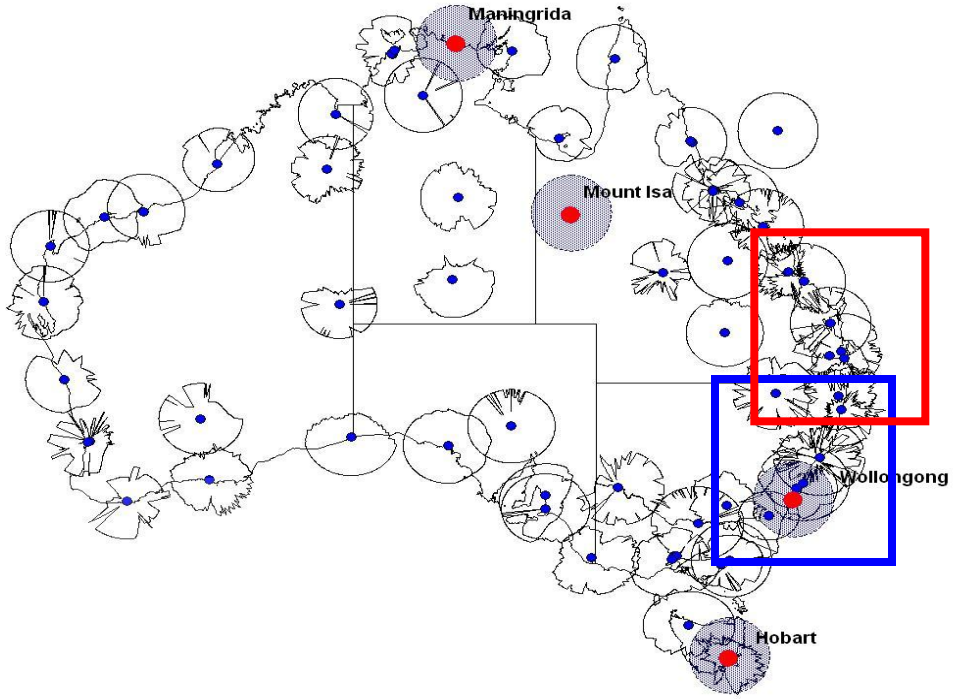
SREP - Strategic Radar Enhancement Project
Next generation City systems

- Higher resolution prediction
  - UM at 1.5km (0.0135º)
  - UM 7.6 soon to be 8.0

- Higher resolution observations
  - Radar wind & cloud/precipitation

- Assimilation
  - 3dVAR at 3km
    - Standard observations
  - Doppler winds
  - Precipitation / clouds
    - Latent heat nudging
    - 1dVAR+3/4dVAR
    - 4dVAR

- More frequent analysis - forecast cycle (RUC)
  - 3-hourly, eventually 1-hourly



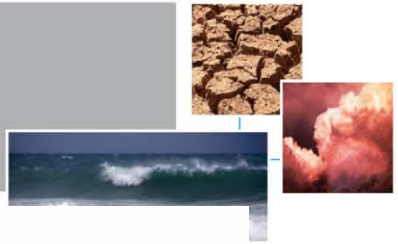
SREP should deliver significant mesoscale NWP upgrades for Australia

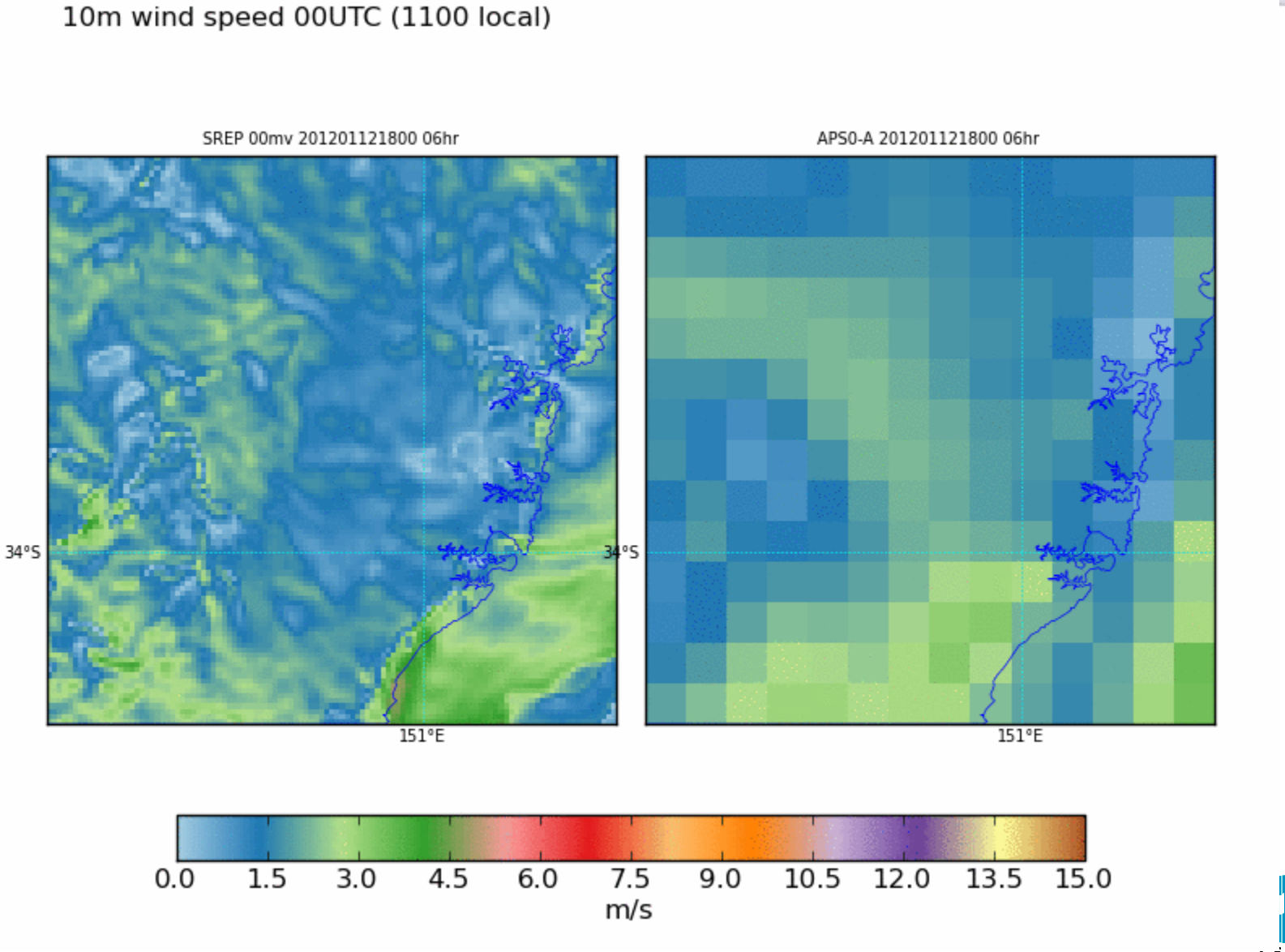
Australian Government
Bureau of Meteorology

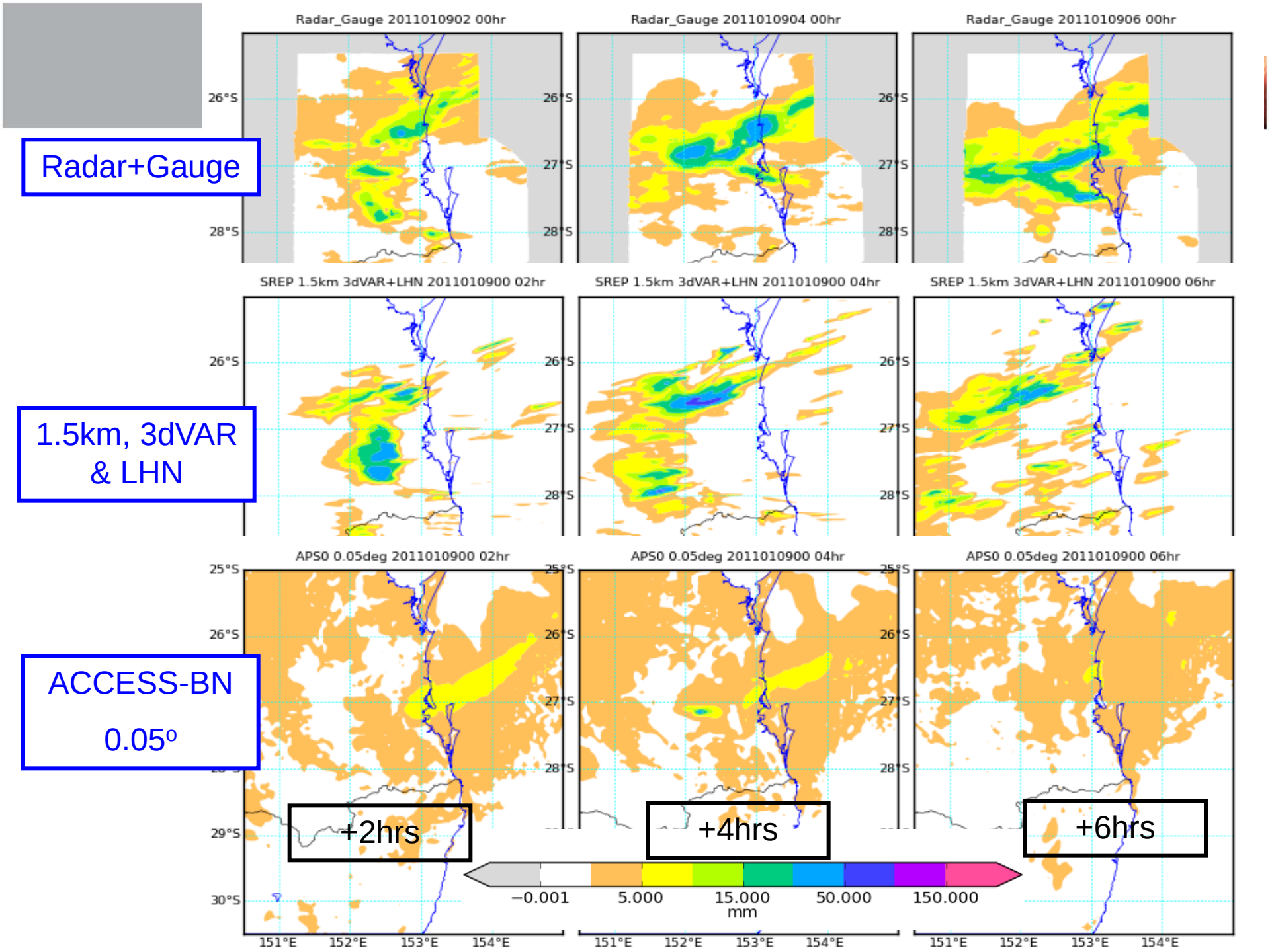
The Centre for Australian Weather and Climate Research
A partnership between CSIRO and the Bureau of Meteorology

CSIRO

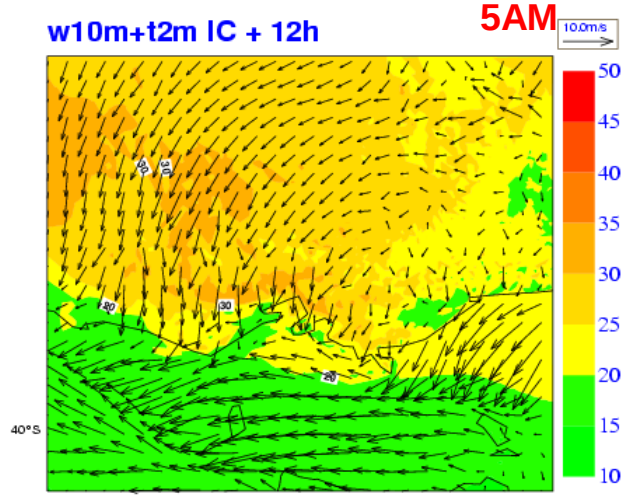
1.5km vs ACCESS-A: 10m wind speed 00Z to 09Z

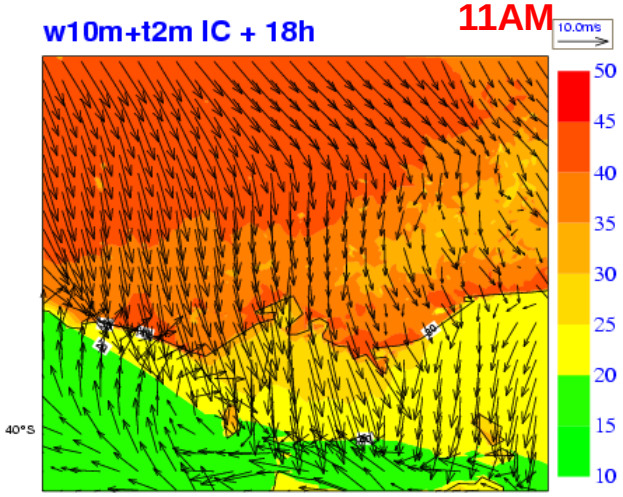




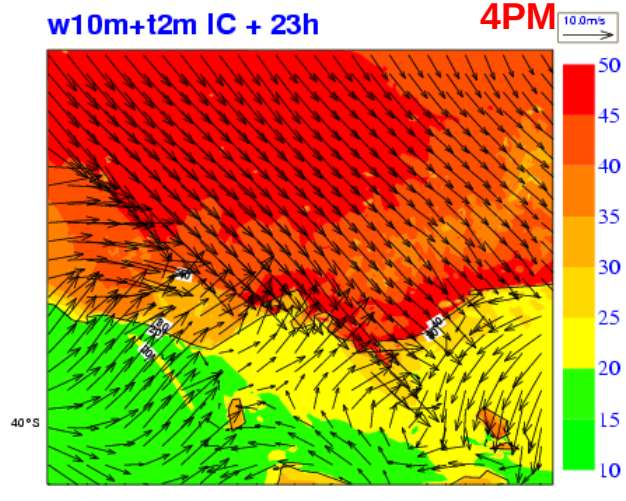

Radar_Gauge 2011010902 00hr
Radar_Gauge 2011010904 00hr
Radar_Gauge 2011010906 00hr
Radar+Gauge
SREP 1.5km 3dVAR+LHN 2011010900 02hr
SREP 1.5km 3dVAR+LHN 2011010900 04hr
SREP 1.5km 3dVAR+LHN 2011010900 06hr
1.5km, 3dVAR & LHN
APS0 0.05deg 2011010900 02hr
APS0 0.05deg 2011010900 04hr
APS0 0.05deg 2011010900 06hr
ACCESS-BN
0.05o
+2hrs
+4hrs
+6hrs
−0.001
5.000
15.000
50.000
150.000
mm
25°S
26°S
27°S
28°S
29°S
30°S
151°E
152°E
153°E
154°E

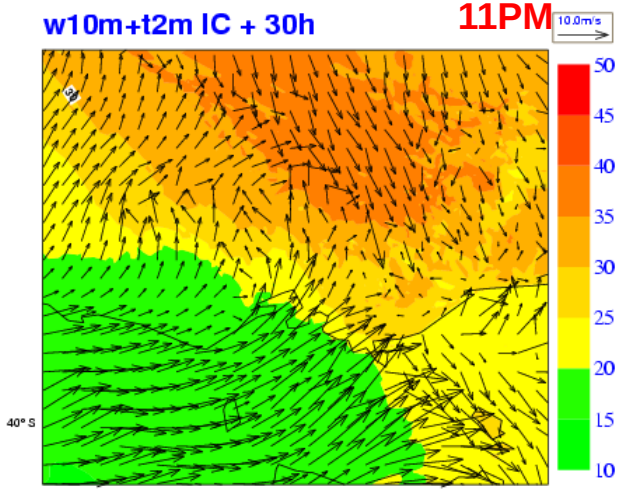
Severe weather prediction





Black Saturday,
7 February 2009,
400 fires,
173 deaths,
2000+ homes lost,
4000+ sq km burnt



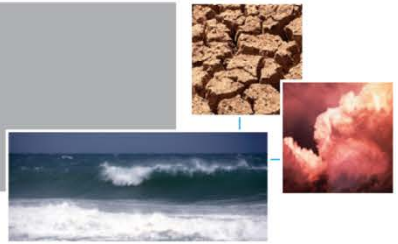


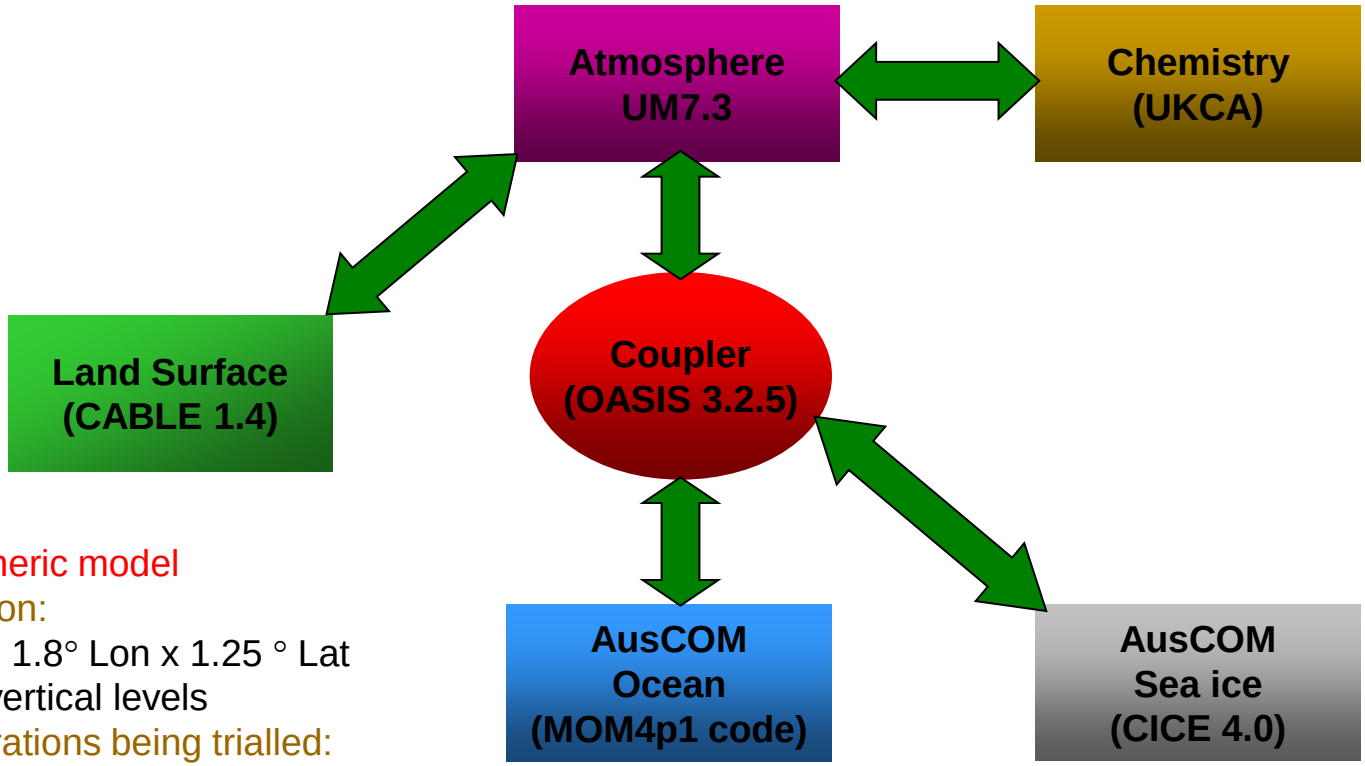
Model run at 3km

Model has been run at 400m resolution

CSIRO

ACCESS CMIP5 Modelling System



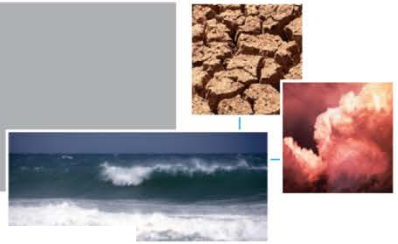


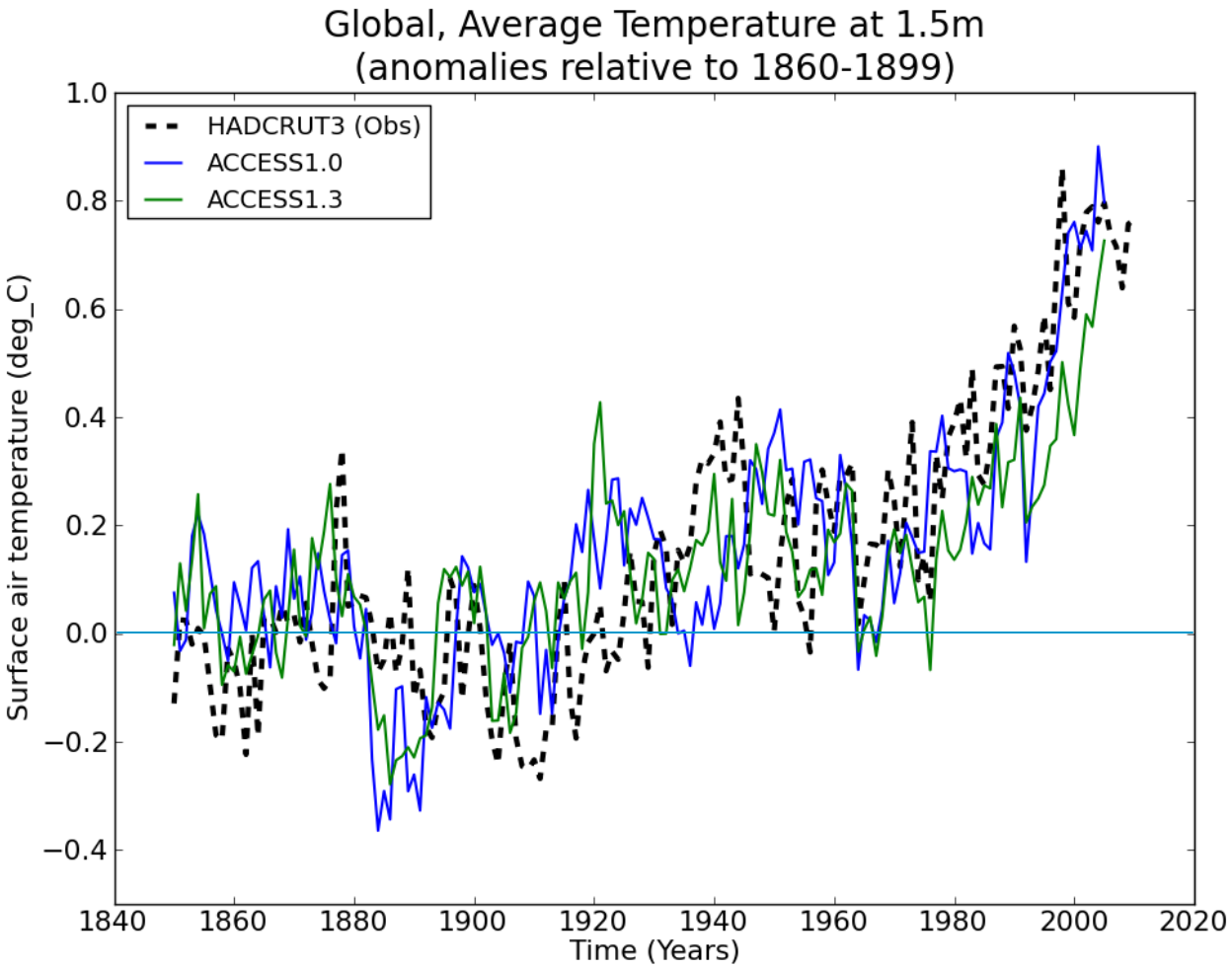
Atmospheric model
Resolution:
N96L38: 1.8° Lon x 1.25 ° Lat
and 38 vertical levels
Configurations being trialled:
HadGEM2
HadGEM2 + PC2 clouds
Proto-HadGEM3

3 hourly flux coupling between models
3.5 simulated years / day

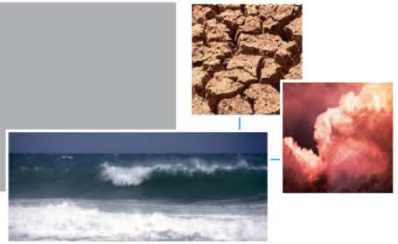
CABLE, AusCOM, CICE, UKCA have been successfully coupled to Unified Model

Annual Mean surface air temperature
Historical Runs

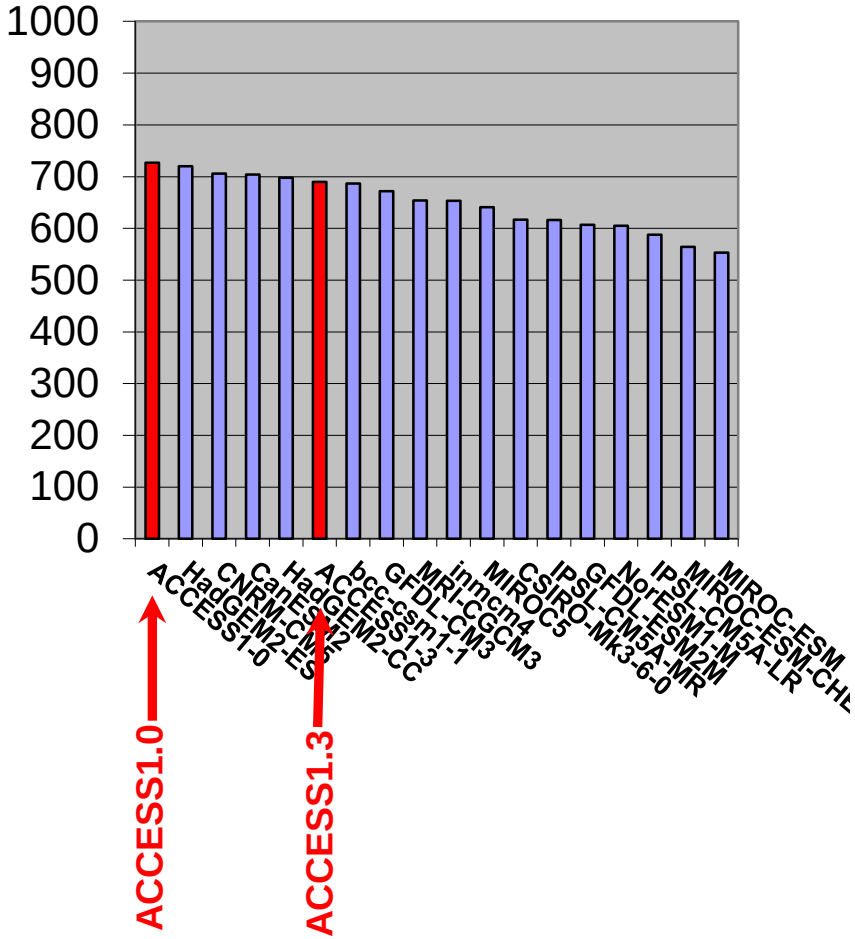




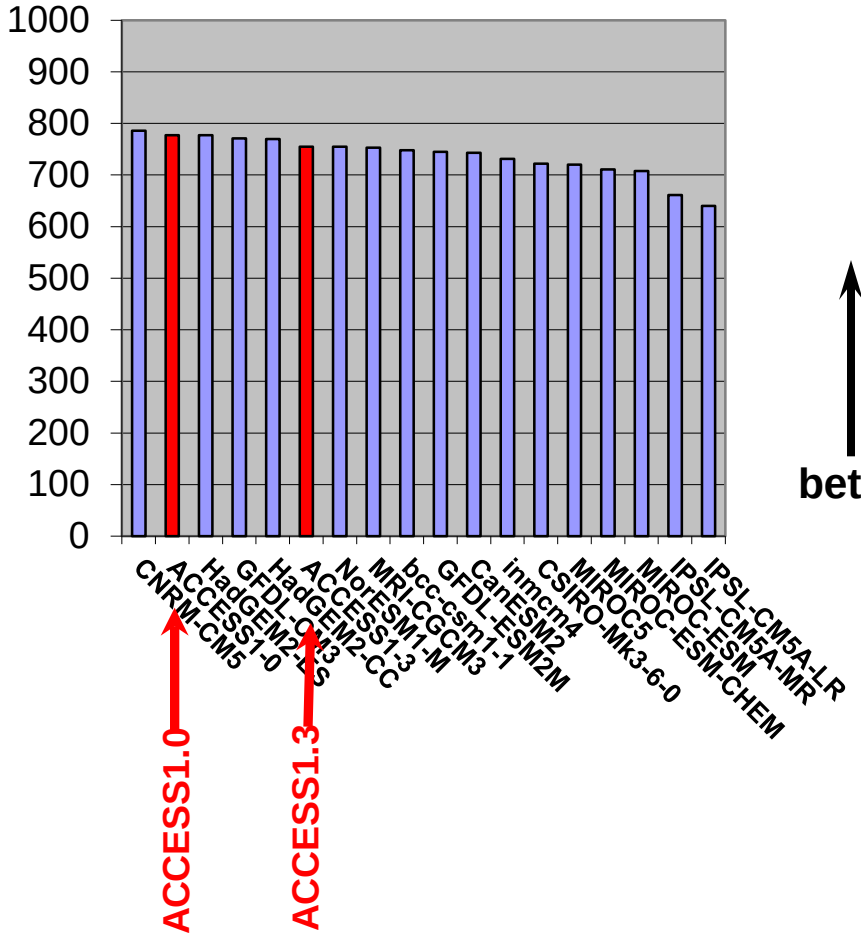
ACCESS coupled model evaluation – ranking versus other CMIP5 models



Average of skill scores (Australia)

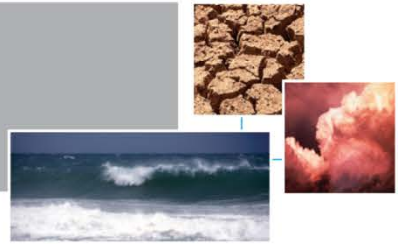


Average of skill scores (globe)



Scores for 3 variables – surface air temperature, precipitation, sea level pressure

Future Directions for NWP



Numerical Weather Prediction

- **Severe weather and high impact weather**
  - High resolution
  - Improved model physics and dynamics
  - SREP
  - **"Model on demand"**
  - Application to downstream systems – hydrology
  - "End-to-end" systems, eg flood forecasting
  - Effectively convey information on severe weather
- **Data assimilation**
  - Hybrid 4DVAR/EnKF
  - New sounders
  - Radar data
- **Multi-week/seasonal/Decadal(?) prediction**
- **Use of coupled models in NWP**
- **Environmental prediction**

Provide outputs that meet societal needs

Quantifying and representing uncertainty

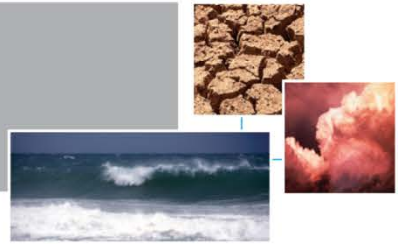
- **Ensembles, including high resolution ensembles**
  - Essential requirement across all areas
  - Optimal strategies for generating perturbations
  - Application to downstream systems
  - Product development

NWP Resolution Plans

	APS1	APS2	APS3	APS4
G	40km L70 4dVAR	25km L70 4dVAR	25km L90 4dVAR	17km L110 Ens-VAR
R & TC	12km L70 4dVAR	12km L70 4dVAR	12km L90 4dVAR Still needed ?	8km L110 Ens-VAR Still needed ?
C	4km L70	4km L70 3dVAR	1.5km(V) L70 3dVAR (R), LHN (3 hourly update)	1.5km(V) L90 3dVAR (R), LHN (1 hourly update)
On-Demand		4km L70 3dVAR	1.5km(V) L70 3dVAR (R), LHN (3 hourly update)	1.5km(V) L90 3dVAR (R), LHN (1 hourly update)
En-G		60km L70 M24	60km L90 M24	35km L90 M24
En-R		24km L70 M24	24km L90 M24	16km L90 M24
En-C				1.5km(V) L90 3dVAR (R), LHN M6

Climate & Weather Science Laboratory

a virtual laboratory for the Australian research community



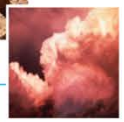
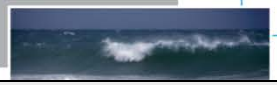
National eResearch Collaboration Tools and Resources (www.nectar.org.au)

Objective:
The virtual laboratory is a new community project to establish an integrated national facility for research in climate and weather simulation and analysis.

Development Organizations:
Australian Bureau of Meteorology (www.bom.gov.au)
Australian National University (facility host) (nci.org.au)
CSIRO Marine and Atmosphere Research (www.csiro.au/cmar)
Centre for Australian Weather and Climate Research (www.cawcr.gov.au)
ARC Centre of Excellence for Climate System Science (www.climatescience.org.au)

Goals:
- To enrich the scientist's access to climate and weather research services,
- To reduce the technical barriers to using state of the art tools,
- To facilitate the sharing of experiments, data and results,
- To reduce the time to conduct scientific research studies, and
- To elevate the collaboration and contributions to the development of the Australian Community Climate Earth-System Simulator (ACCESS)

NeCTAR schematic

Web site

Climate & Weather Science and Simulation Laboratory

Portal Service

- ACCESS Simulation Service
- Model Analysis Service
- Community Data Portal Service
- Future Research Services

Framework & Toolkits

- ACCESS Framework
- Model Analysis Framework
- Data Services Framework
- Service Framework
- Modelling Toolkit
- Model Analysis Toolkit
- Data Toolkit
- Service Tools & Methods

Core API's

Core Data, Analysis, and Graphics Layer

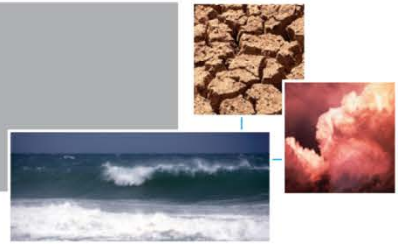
Collab Service

Facility Interfaces and Data Library Services

National Facilities

- NCI Storage and Compute Facility
- RDSI Storage Facility
- IMOS & TERN Facilities
- RDSI & ESG nodes
- …

Present TeraScale Systems

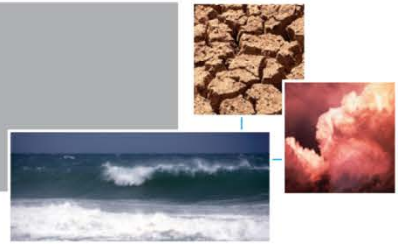


- Bureau and ANU/NCI procured systems in 2009 from Sun Microsystems (now Oracle)

  - **Bureau of Meteorology's HPC system called "Solar"**
    - **50 Teraflops**
    - **115 Terabytes of SAS disk storage**
    - **576 nodes, each comprising 2 Xeon (Nehalem) quad core chips**
    - **Centos 5.3 Linux, SGE 6.2u6 and Lustre 1.8.4**
    - **Primarily NWP and Seasonal systems run**

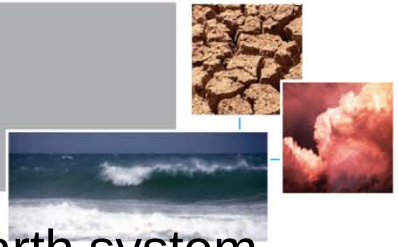
  - **ANU/NCI HPC system called "Vayu"**
    - **140 Teraflops**
    - **1492 nodes, each comprising 2 Xeon (Nehalem) quad core chips**
    - **800 Terabytes of SATA disk storage**
    - **Centos 5.6 Linux, OpenPBS and Lustre 1.8.7/8**
    - **CSIRO has a 24% share in the ANU/NCI HPC system**
    - **Climate modeling is primarily run at ANU/NCI**

2013 BoM HPC upgrade and future HPC



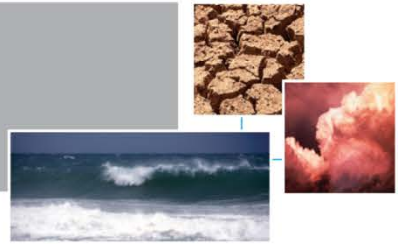
- Refresh of the current supercomputer due in 2013. New system based on Intel Sandy Bridge chips, capable of 100-150 Teraflops, 200+ Terabytes of storage, sited in new data centre. Details are still under negotiation with Oracle.
- Next generation HPC will arrive around 2015/16. Tender commencement is dependent on Australian Government funding.
- Current refresh and next generation HPC will be focused on Operational systems.
- BoM/CAWCR research is moving to a new ANU/NCI supercomputer at start of 2013.
- Collaborations on aspects of HPC with Fujitsu and Intel, running ACCESS systems at ANU/NCI, are in negotiation.

2013 Petascale HPC Facility at ANU/NCI



- New $50m Petascale HPC facility for climate change, earth system science and national water management research, due Jan 2013.
- Centre piece is 1.2 Petaflop supercomputer from Fujitsu comprising 57,472 cores on 3592 nodes (Intel Sandy Bridge 8-core), 158 Terabytes of memory, 9 Petabytes of SAS disk storage, running Centos Linux, OpenPBS and Lustre.
    - 255 TB/s memory bandwidth across system
    - 71GB/s memory bandwidth on node
    - 2.8GB/s MPI BW per direction per node
    - 333 Gflops peak performance on node
    - 54.5 Gbit/s network bandwidth
    - 120GB/s filesystem performance
    - 4200 SAS drives in disk farm
    - 1.5MW of power use

UM Performance on Sandy Bridge Intel Processors



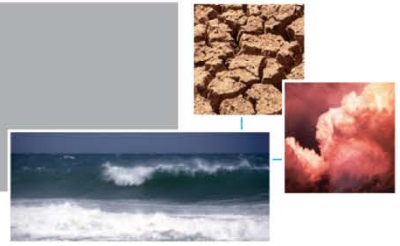
1. Intel Processor Reference Comparison

- Intel Nehalem EP and Sandy Bridge EP processors

2. Unified Model Benchmarks

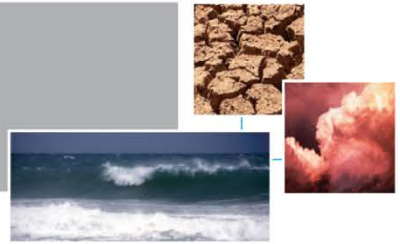
- Intel Nehalem EP and Sandy Bridge EP processors
- N48L38 benchmark
  - MPI scalability
- N512L70 benchmark
  - I/O server scalability
  - Overall scalability and performance

Intel Compute Nodes



	Nehalem EP node X5570	Sandy Bridge EP node E5 2670	Comparison
Processor cores	2 x (2.93 Ghz, 4-core)	2 x (2.6 Ghz, 8-core)	2.0x cores
Memory	3 x DDR3-1333 Mhz	4 x DDR3-1600 Mhz	33% more DIMMS
Peak Performance	94 GF	333 GF	3.5x
SPECfp_rate/node (SPECfp_rate/core)	168 (21.0)	445 (27.8)	2.6x (1.3x)
Memory Bandwidth STREAM Triad (full node)	36 GB/s	71 GB/s	~2.0x
CPU Power	95 W	115 W	21% increase
Node Power	368 W	366 W	same

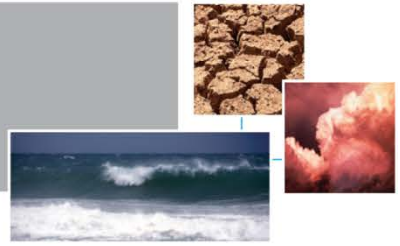
Unified Model Benchmarks



- Unified Model v8.0 is testing on Intel Sandy Bridge platforms.
  - UM N48L70 tested for single node performance
  - UM N512L70 tested for multi-node performance
  - Collaboration with Intel/HP and Enda O'Brien for UM N512L70 benchmark runs on a Sandy Bridge system.

- Result from UM N48L70 runs:
  - Intel Sandy Bridge compute node can execute the benchmark twice as fast as an existing Intel Nehalem compute node and with identical power consumption. (2x increase in flops/watt)
  - MPI process placement is critical for good performance.

Unified Model N48L70 Benchmark

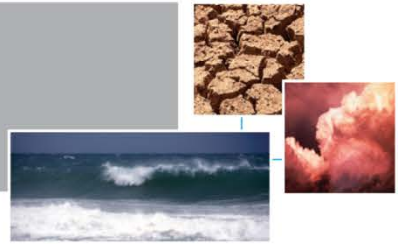


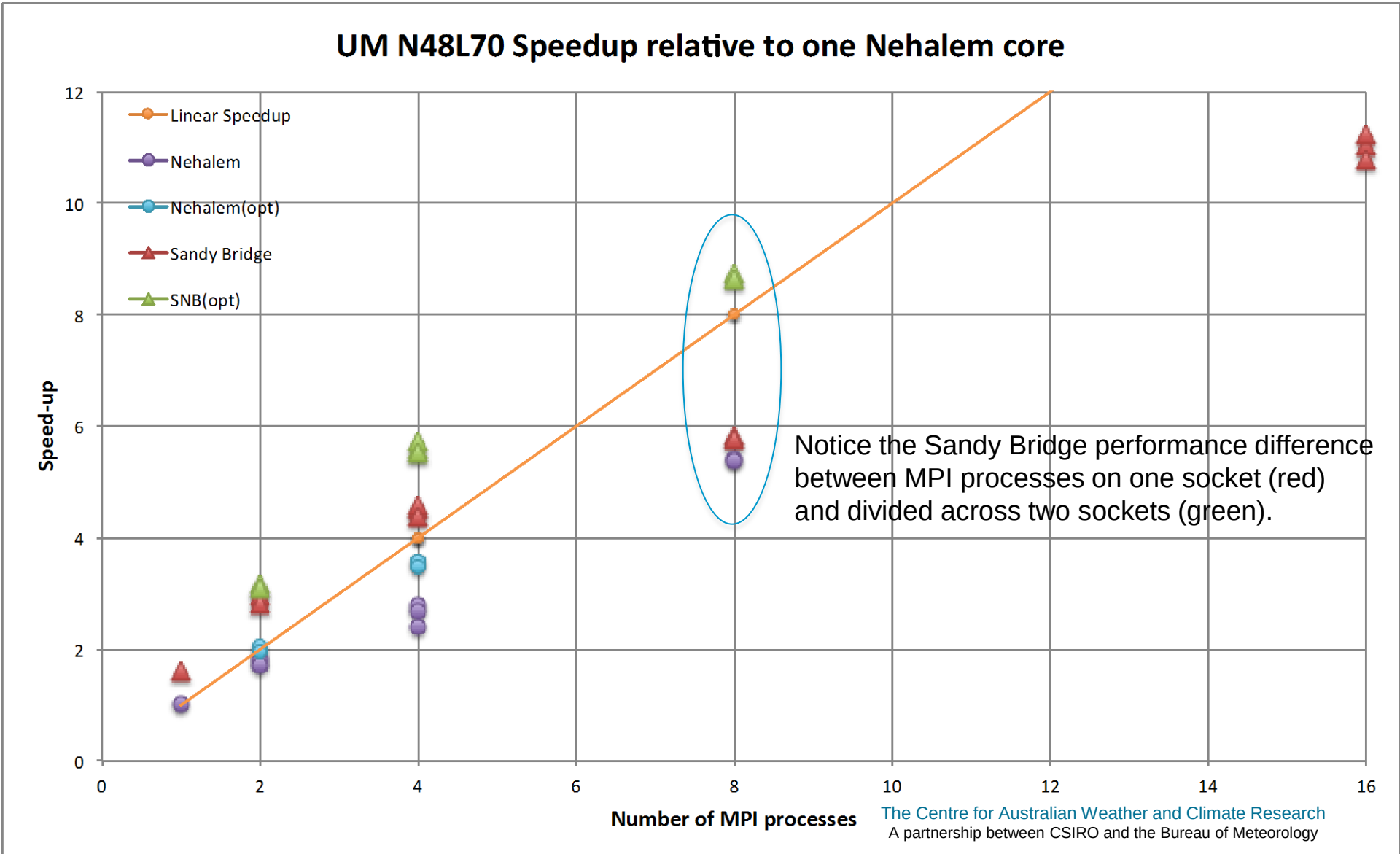
- Unified Model N48L70 benchmark runs on a single compute node
  - Based on UM v8.0 code
  - I/O server is turned off
  - MPI only parallel run, 1 thread per MPI task
  - N48 model run on a dual socket Intel compute node
    - Dual 2.6 Ghz, 8-core Intel Sandy Bridge EP processors with DDR3-1600 memory
    - Dual 2.93 Ghz, 4-core Intel Nehalem EP processors with DDR3-1333 memory
    - Speedup relative to single Nehalem core
    - For UM N48, Intel Sandy Bridge node is twice as fast as Intel Nehalem node

	Cores		Elapse Time		Speedup (Nehalem)		NHM vs SNB
Domain	Nehalem	Sandy Bridge	Nehalem	Sandy Bridge	Nehalem	Sandy Bridge	Difference
Single Core	1	1	3830	2410	1.00	1.59	1.59
Equal Cores per Socket	4	4	1376	839	2.78	4.56	1.64
Single Socket	4	8	1376	657	2.78	5.83	2.09
Optimal using Dual Socket	4	8	1072	439	3.57	8.72	2.44
Dual Socket	8	16	707	341	5.42	11.23	2.07

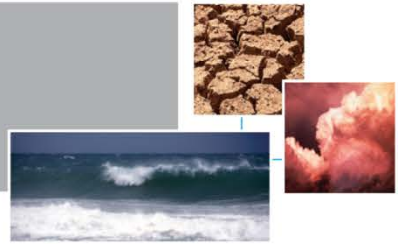
intel compiler 12.1.8.273
with : -g -i8 -r8 -O3 -xHost -fp-model precise

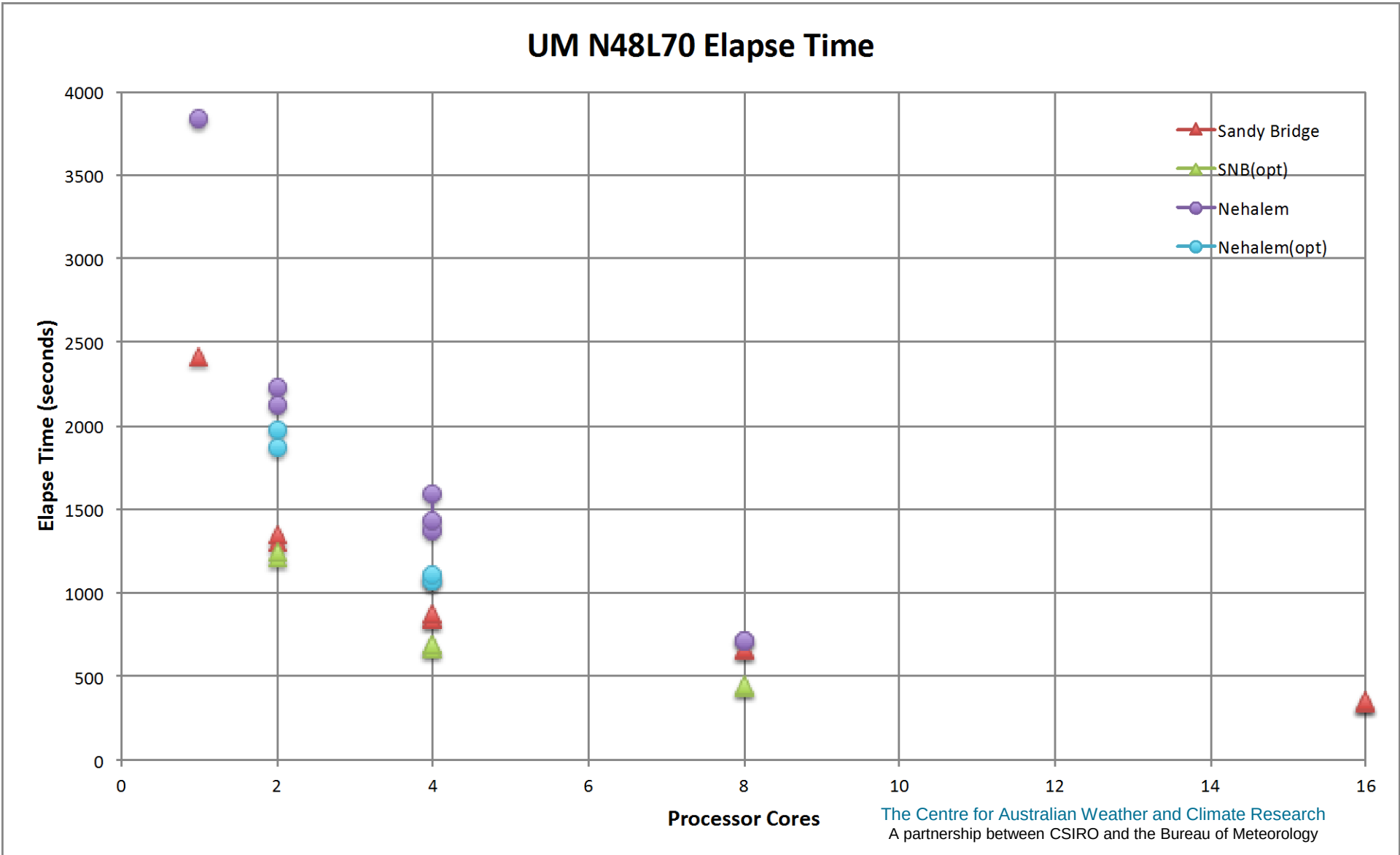
Unified Model N48L70 – Single Node



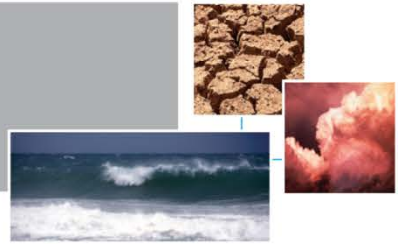


Unified Model N48L70 – Single Node





N512L70 Benchmark Results

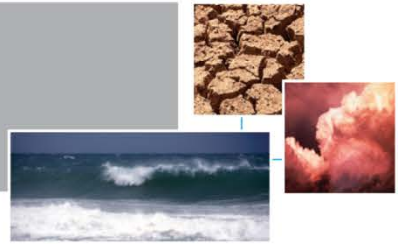


Improvements in application performance due to faster memory

Table 2

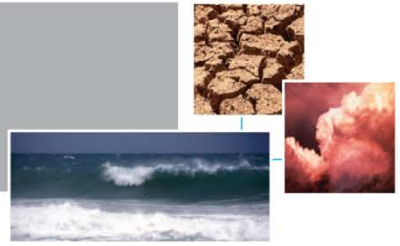
UM 1-day Forecast (run-times in s)

Unified Model N512L70 Benchmark



- Main results to-date:
  - All the main code loops vectorize to use the SandyBridge AVX vector instructions.
    - But performance is limited more by the ability of the memory system to deliver data to the CPUs than by the processing power of the CPUs themselves.
  - Correct process-placement and thread affinity at run-time is critical for good performance on SandyBridge.
    - Further placement and affinity subtleties are introduced when HyperThreading is enabled on the nodes.
    - Memory affinity is necessary on Linux systems to control runtime variability (ANU/NCI)
  - The most promising approach to improve overall performance is to dedicate some processes to I/O only.
    - Since a large fraction of total run-time is taken up by I/O.
    - There is little or no potential in code re-structuring to improve SandyBridge performance.

Unified Model N512L70 Benchmark



- Main results to-date (continued):
  - On a “core-by-core” comparison, performance of UM on Intel SandyBridge is approx. 10% faster than on the older Nehalem processors (assuming similar clock-speeds).
    - That 10% is due mainly to the AVX vector instructions.
  - On a “node-by-node” comparison, a SandyBridge cluster is more than 2 times faster than a Nehalem one by virtue of having twice as many cores per node.
  - For the problem size provided (N512L70), the target of 24-hr simulation in less than 500s can easily be met in the no-IO case, but not in the full-IO case.
    - For the full-IO case, the best result obtained was 856s, using 32x32 decomposition on 1024 cores (64 nodes) and with no multi-threading (i.e., no OpenMP).

Australian Government
Bureau of Meteorology

The Centre for Australian Weather and Climate Research
A partnership between CSIRO and the Bureau of Meteorology

CSIRO

Robin Bowen
Senior Information Technology Officer
Earth System Modelling Programme
email: r.bowen@bom.gov.au

Thanks to my CAWCR Colleagues and to our partners in HPC:
ANU/NCI, CSIRO, Intel, Oracle, and the UK Met Office

Thank you

www.cawcr.gov.au