Fast time series analysis of wave hindcast data

Robert Davy [1], Ron Hoeke [2], Claire Trenham [2], Julian O'Grady [2], Mark Hemer [2], Rebecca Gregory [2], Kathleen McInnes [2]

[1] CSIRO Information Management & Technology, Canberra; [2] CSIRO Oceans and Atmosphere, Aspendale.

www.csiro.au

An IMT eResearch Collaboration Project initiated by the Sea Level, Waves and Coastal Extremes team.

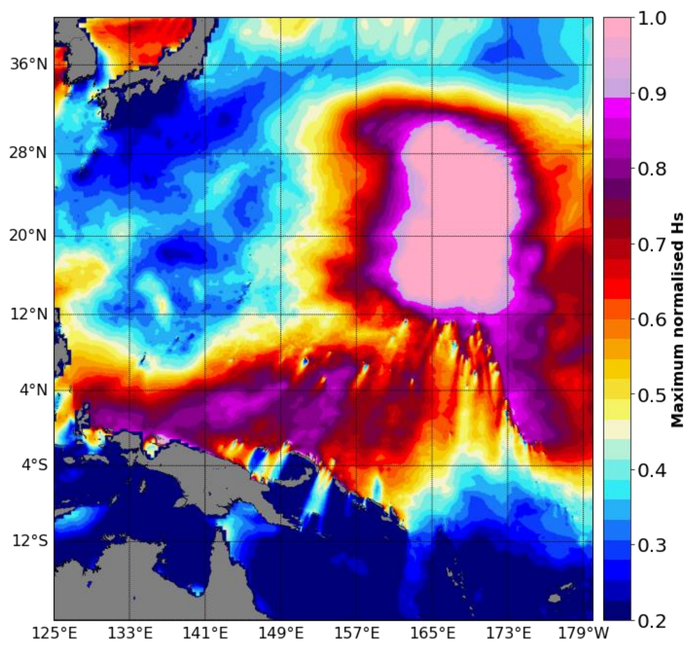
Introduction

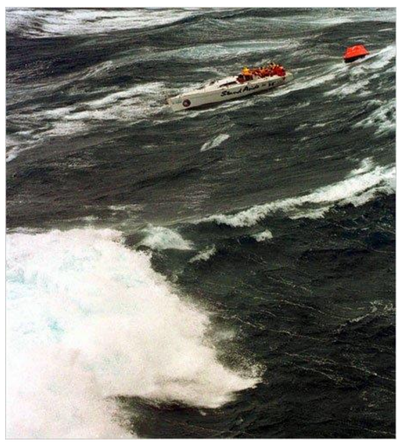
- Through a CSIRO – Bureau of Meteorology partnership, a spectral wind-wave hindcast model has been run to provide gridded wave statistics and directional wave spectra at hourly intervals. This highly cited dataset [1,2] is recognised as the highest spatial and temporal resolution wave hindcast available for the Australian and Pacific Regions.
- The wave hindcast output is optimised for spatial analyses. Constructing a 30+ year hourly time series at a grid point can take around 90 minutes, as data has to be extracted from many files to cover the temporal range. Large scale analysis of time series data is not practical.

An eResearch Collaboration Project was initiated with the aim of optimising rapid access to this data for time series, particularly extreme value analysis.

Method

- Data has been reorganised into spatial tiles, as follows:
  - extract data by tile,
  - concatenate in time, then
  - apply NetCDF chunking mainly in the time dimension.
- ChunkSizes (time, lat, lon) = (35064, 1, 15)
- Due to the large memory requirements, processing is performed using bash/python scripts on CSIRO's large memory multiprocessor (Ruby).
- There is capability to update tiles as new data comes in (re-chunking is required).





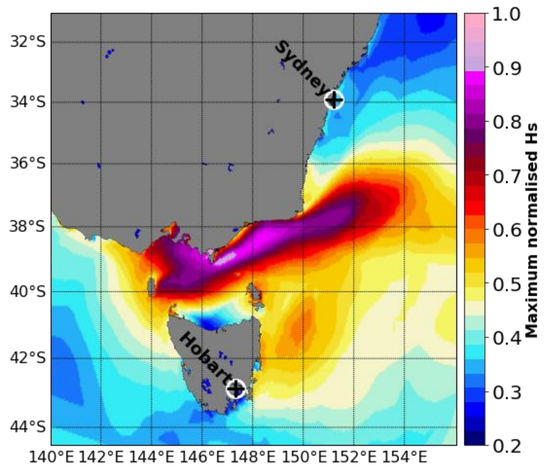


Figure 2: Examples of extreme storm wave event analysis for a mid-latitude storm which caused widespread flooding of Pacific Islands in 2008 [3] (left map and pictures) and during the 1998 Sydney to Hobart Yacht Race, during which five yachts were lost, 7 others abandoned, six sailors lost their lives and 55 sailors had to be rescued from their yachts by ships and helicopters [4] (right map and pictures). Maps represent the maximum significant wave height (Hs) over the duration of the event (several days), normalised by the 10-year ARI, i.e. map colours greater than 1 indicates storm wave heights greater than the 1-in-10 year event.

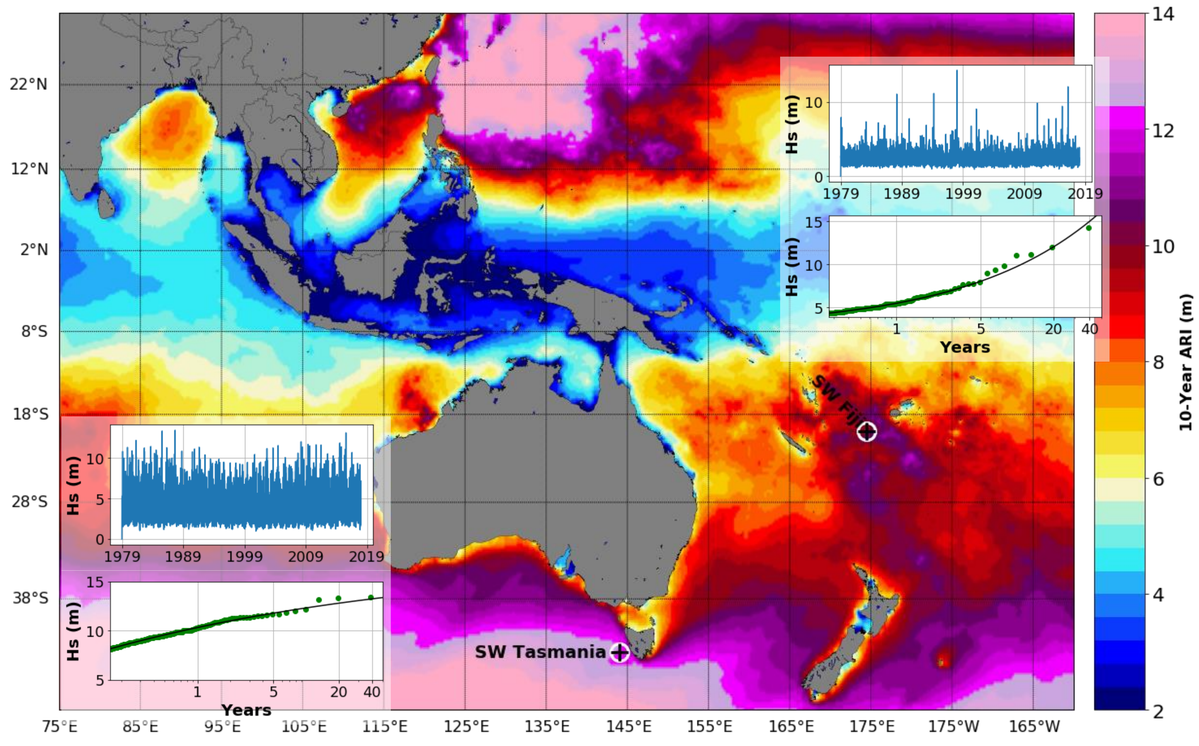


Figure 1: Example 10-year average return interval (ARI) significant wave height (Hs), calculated from a generalized pareto (GPD) fit of all hourly data for years 1979 – 2018 tiles in the map region. Two example locations, south-west (SW) Tasmania and south-west (SW) Fiji archipelago illustrate the full hourly timeseries (upper plots) and empirical and fitted ARI values (green circles and black lines, respectively, in lower plots), both sites are plotted on the same scale.

Processing Results

Initial creation of the tile set takes a few days in total using 10 cores, each with 12.8 GB RAM, on Ruby. Thereafter, time series analysis is much faster:

- Retrieval of the hourly time series at any grid point (>340k time steps) now takes around 0.1 second, which represents a speedup of several orders of magnitude.
- There is a trade off between access time and tile size – smaller tiles are generally faster to work with.
- Extreme value analysis of the entire Australian coastal domain can be done on the Pearcey cluster (using job parallelism) in around 10 minutes (25 tiles @ 0.067° spatial resolution).
- The entire global ice-free ocean area can be processed at lower spatial resolution (60 tiles @ 0.4°) in around 20 minutes. See Table 1.

Table 1: Processing time of full datasets after tiling and netCDF chunking.

DOMAIN	# TILES	# OCEAN GRID POINTS	PROCESSING TIME BEFORE TILING	PROCESSING TIME AFTER TILING
1 point	-	1	90 minutes	0.1 seconds
glob_24m	60	196k	Not possible	20 minutes
pac_4m	31	138k	Not possible	24 minutes
aus_4m	25	67k	Not possible	10 minutes

Conclusion

"Even with SSD, rechunking data takes a relatively long time. Is it ever worth it? I think it is, for important datasets that will be written once but read many times." [5]

Restructuring the gridded wind-wave hindcast data has dramatically improved computational speed for time series analysis of the data. This increase in speed has made possible:

- Rapid assessment of the significance of extreme ocean wave events in a climatological context.
- Statistical-significance based forecast warnings of future extremes.
- Ease of access to this data also improved by this process.

FOR FURTHER INFORMATION

Robert Davy

e robert.davy@csiro.au

w confluence.csiro.au/display/SC

REFERENCES

[1] Durrant, Greenslade, Hemer, Trenham, (2014). Global Wave Hindcast focussed on the Central and South Pacific. CAWCR Tech. Rep. 70, 1–54.

[2] Hemer, et al. (2016). A revised assessment of Australia's national wave energy resource. Renewable Energy doi:10.1016/j.renene.2016.08.039

[3] Hoeke, et al. (2013). Widespread inundation of Pacific islands triggered by distant-source wind-waves. Glob. Planet. Change 108, 128–138.

[4] Greenslade, (2001). A wave modelling study of the 1998 Sydney to Hobart yacht race. Australian Meteorological Magazine 50, no. 1.

[5] Russ Rew (2013). Chunking Data: Choosing Shapes. https://www.unidata.ucar.edu/blogs/developer/en/entry/chunking_data_choosing_shapes

ACKNOWLEDGEMENTS

In addition to support from **CSIRO's eResearch** Programme, aspects of this study are supported by The Earth System and Climate Change Hub of Australian Government's National Environmental Science Programme (**NESP**). We also acknowledge the CSIRO High Performance Computing facility which provided computational support for the work described in this study.