PO.063

The use of zarr files to store and study 4D wind flow BLOCKS

Marta Gil Bardají
Vortex Factoria de Càlculs

Abstract

Current atmospheric modeling is technically capable of generating historical wind data time series over a region for several heights. This 4-dimensional gridded time series data set is what we internally name BLOCKS at Vortex and will allow wind engineers to study the wind flow in great detail and use it as input for complex calibration, wakes and production studies.

Besides the computational challenges of performing such a simulation for high resolutions and long periods of time, another limitation is the storage and handling of the data set itself. BLOCKS are so voluminous and complex that traditional wind software formats are inadequate to deal with them (yet) and this hinders their widespread. Fortunately, there exist interesting candidates to become the next industry standard and in this presentation we will show why Vortex has chosen one of them: the **Zarr files**.

Zarr is a cloud-friendly data format implemented in Python that stores chunked, compressed N-dimensional arrays [1]. Zarr is a young project but already very popular especially in parallel computing and cloud storage contexts. A key feature of a zarr file is that the data arrays are divided into chunks (pieces) and each chunk is compressed. The optimal chunk shape depends on how one will access the data and the performance can vary greatly if chunks are chosen differently. Therefore, choosing the correct chunking for the data is the essential decision to create the "best" zarr possible.

Objectives

This study aims to show examples of 4D simulations stored in different formats to give perspective of expected storage space and data retrieval times. The aim of this presentation is to encourage the use of zarr files by providing basic guidelines for chunking, since we envision the zarr format as the new standard for dealing with BLOCKS 4D datasets in the industry.

Methods

The Data Set consists of 1080 netCDF files with daily data (48 registers per day) for a 300x300 grid and 9 vertical levels containing temperature (T) and wind components (U and V) data corresponding to WRF (Weather Research and Forecast) simulations.

From this input data, collections of netCDFs with fewer days and fewer spatial dimensions are generated. They can be opened in python (using xarray package with dask enabled) and we could perform some operations like:

- Compute a mean temperature layer (T) at a level provided by the data set.
- Compute a mean wind speed layer (using U and V) at an interpolated level.
- Extract a time series at coordinates and height provided by the data set.
- Extract a time series by interpolating its coordinates and height.

The goal is to study how do the period and the spatial dimensions of the data set influence these times. Moreover, a data set in python can be chunked to allow parallel computation using dask (this behavior will strongly depend on the machine used). Different chunking strategies will be compared and the chunks will be used to store the data sets in zarr format. The same operations performed over the zarr files instead of over the daily netCDF files will prove to be very efficient.

Chunking Strategies

By default, opening the daily netCDF files in xarray creates a data set with chunks of length 48 in the time dimension: that is the 'day' chunks strategy. Another chunking method is to increase the chunking in time using chunks of a length of approximately 1 month ('month') . Next step is to chunk the levels ('month_lev') and then to make spatial chunks of 20x20 ('month_sp20'). The last strategy is to maintain the spatial chunks and increase the time chunks to contain a year worth of data ('year_sp20').

	Time	Latitude	Longitude	Level
day	48	No	No	No
month	48x30	No	No	No
month_lev	48x30	No	No	1
month_sp20	48x30	20	20	1
year_sp20	48x366	20	20	1

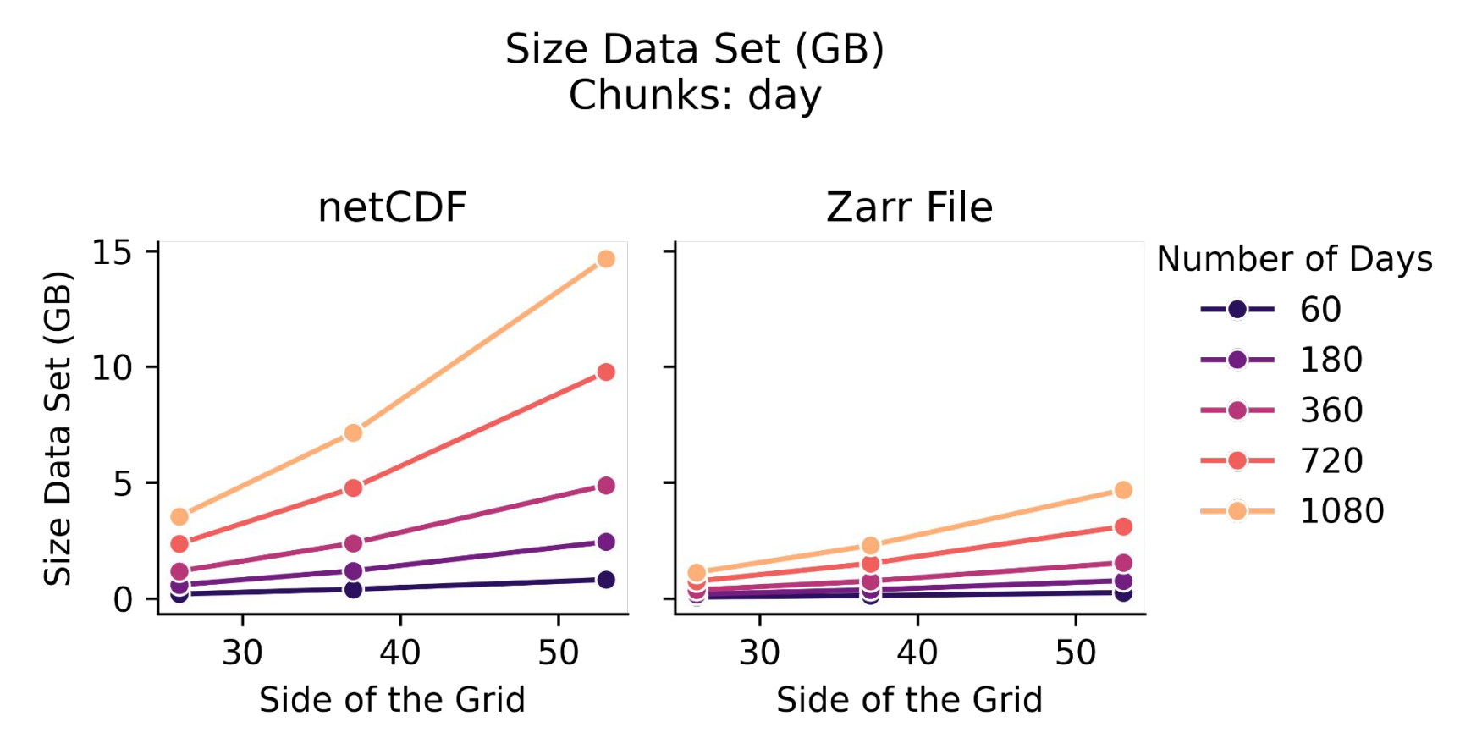
Complete results

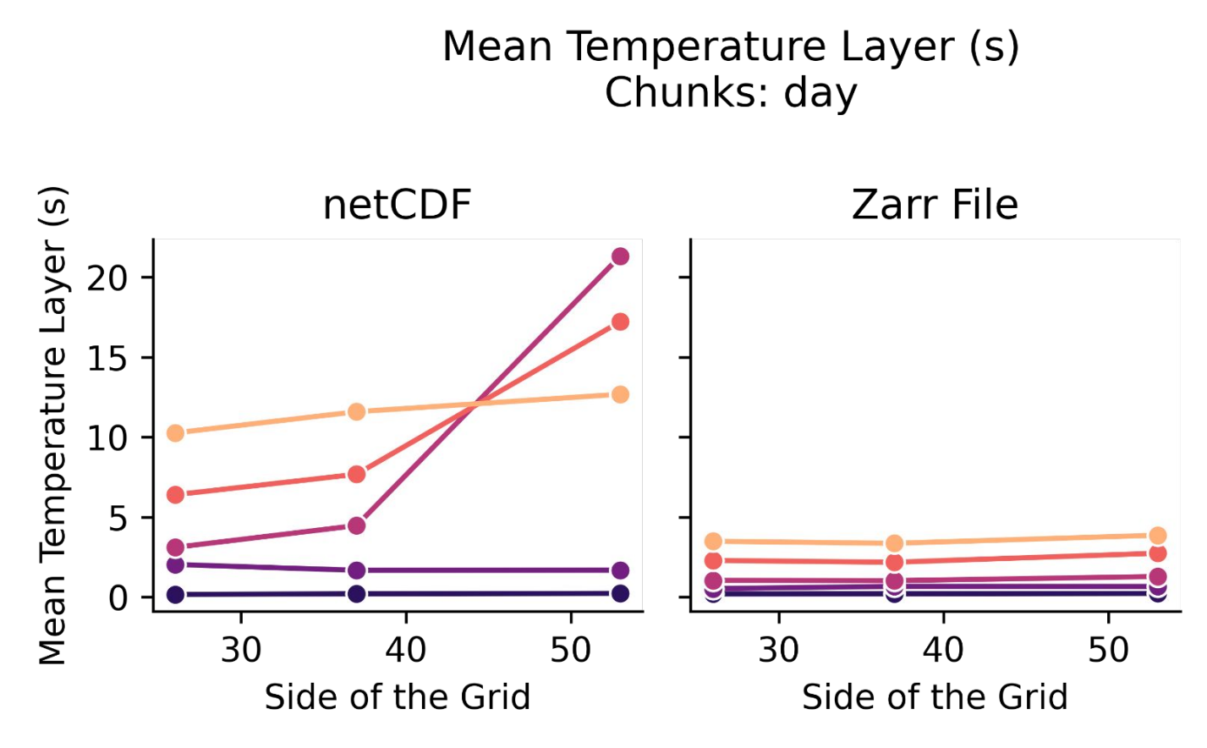
In this poster only a summary of the results and conclusions can be included. If the reader is interested, please contact me (marta.gil@vortexfdc.com) to obtain the full results and code specifications. The numbers showed in this poster will depend on the computer and software versions used, and should therefore be interpreted as a reference comparison.

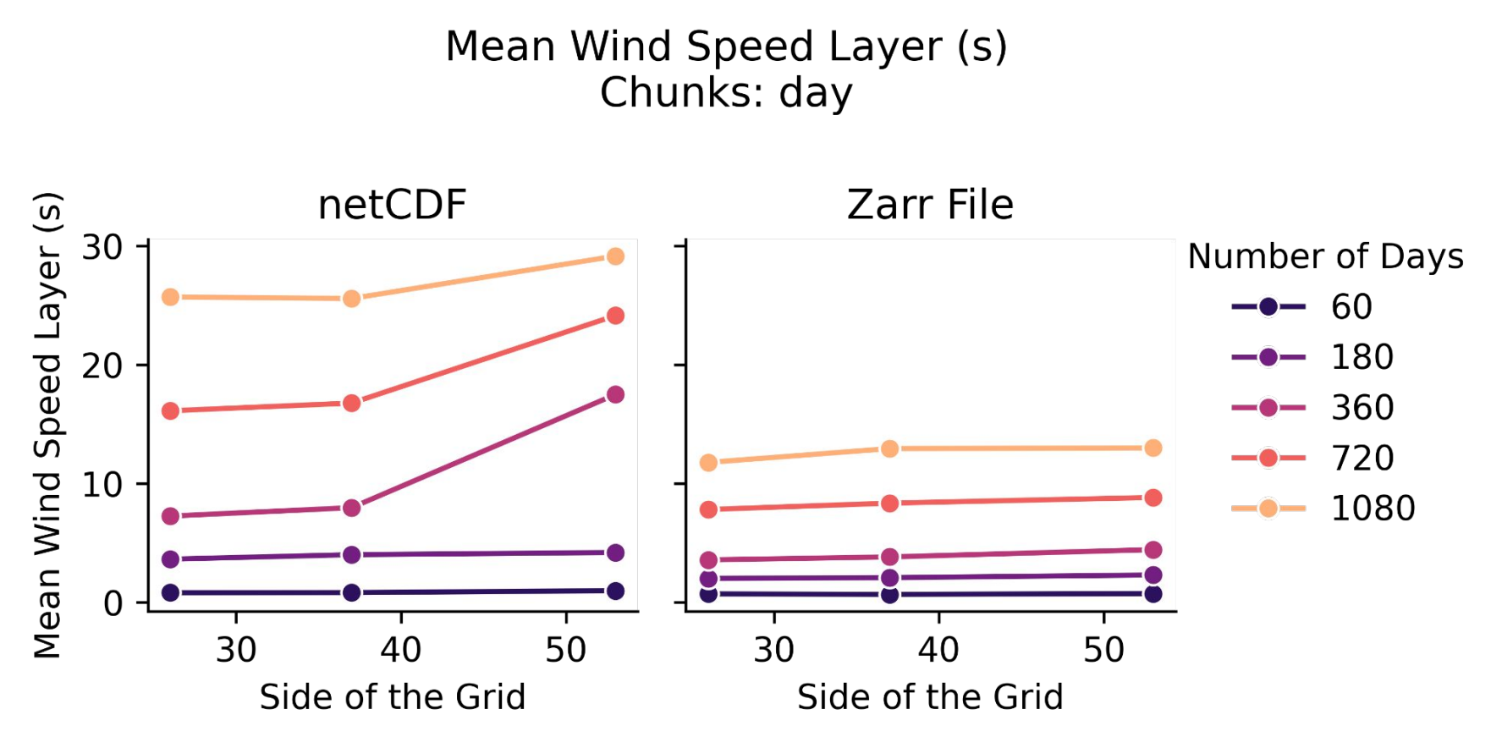
Results

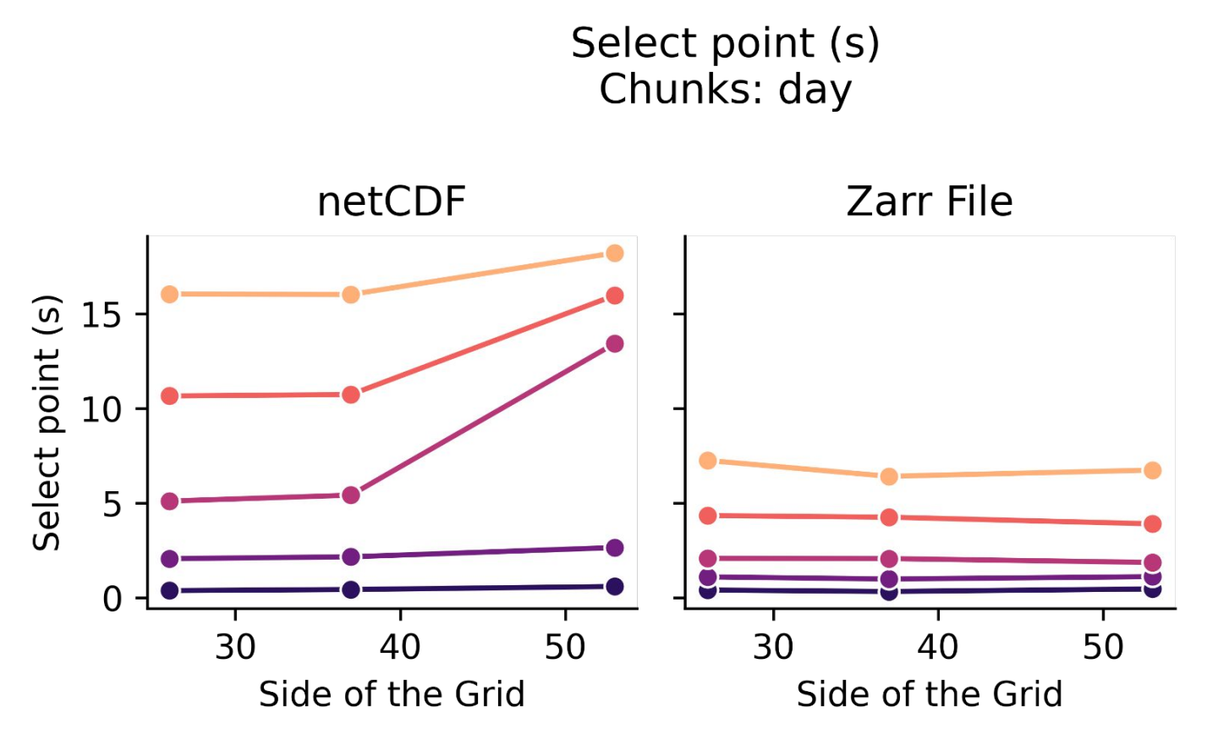
netCDF vs Zarr Format

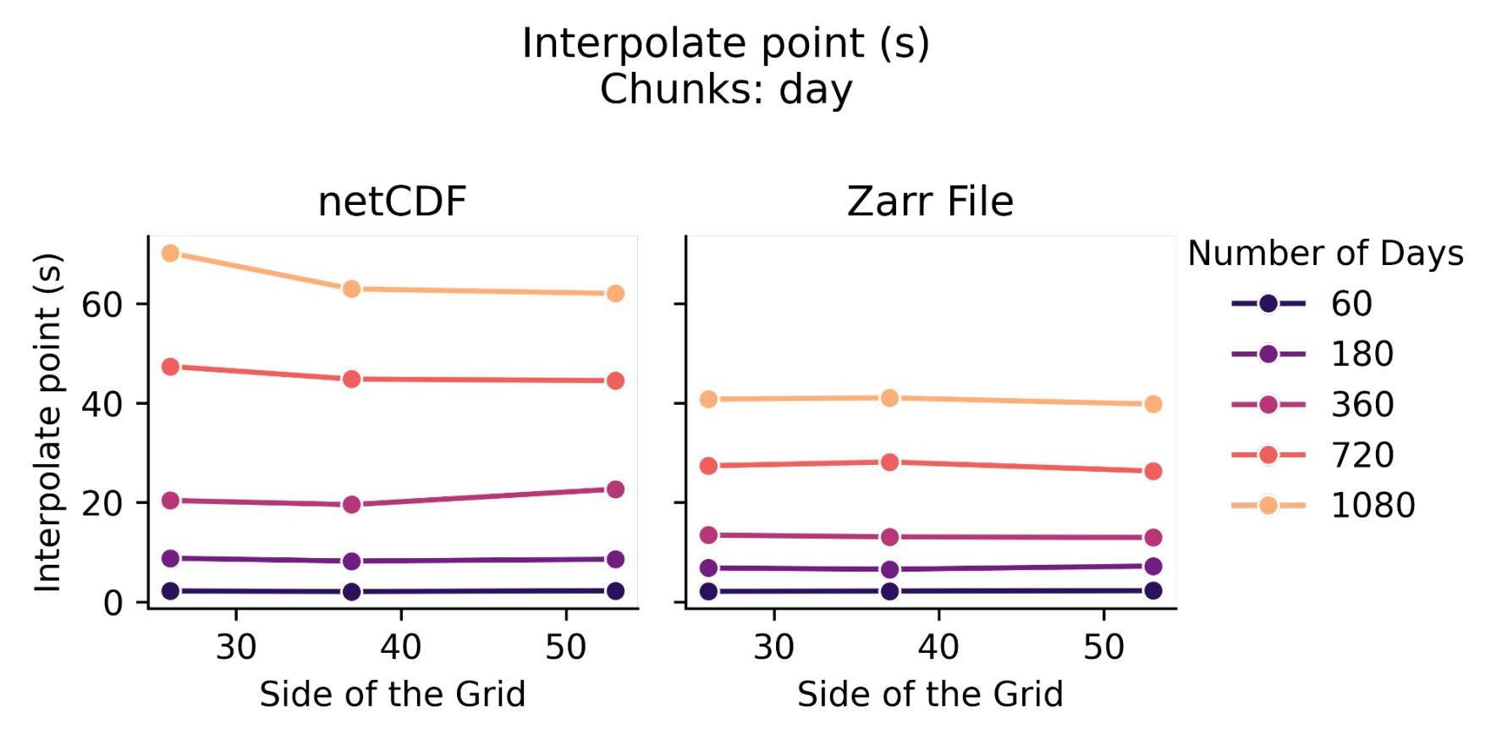
This section compares the size and performance of netCDF and Zarr Files data sets. The default 'day' chunking is used in both formats. Zarr files perform better and seem to scale better than netCDF files.





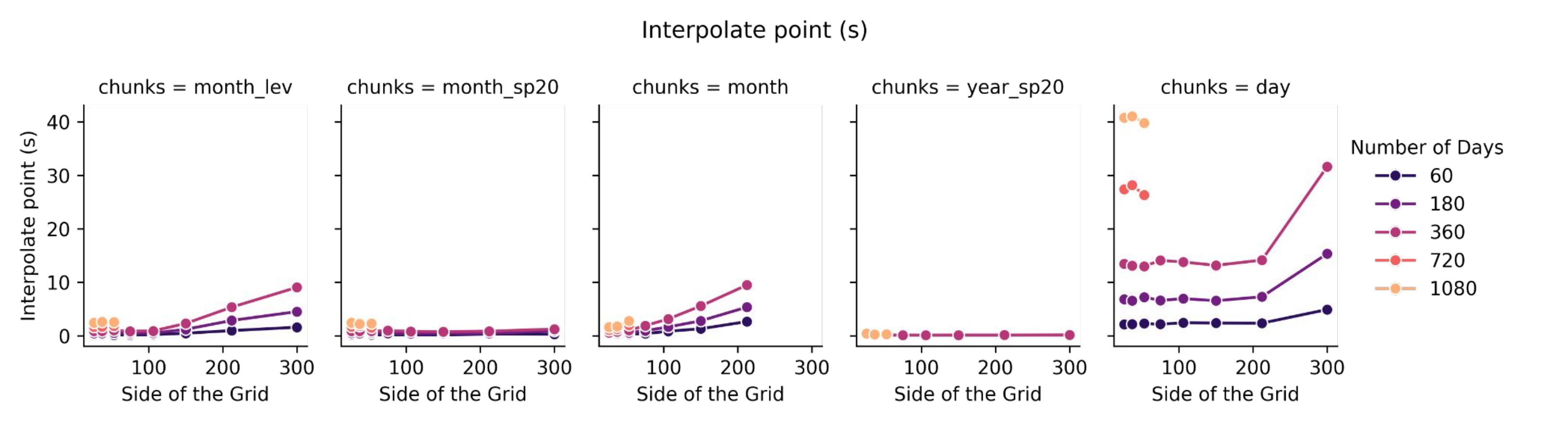


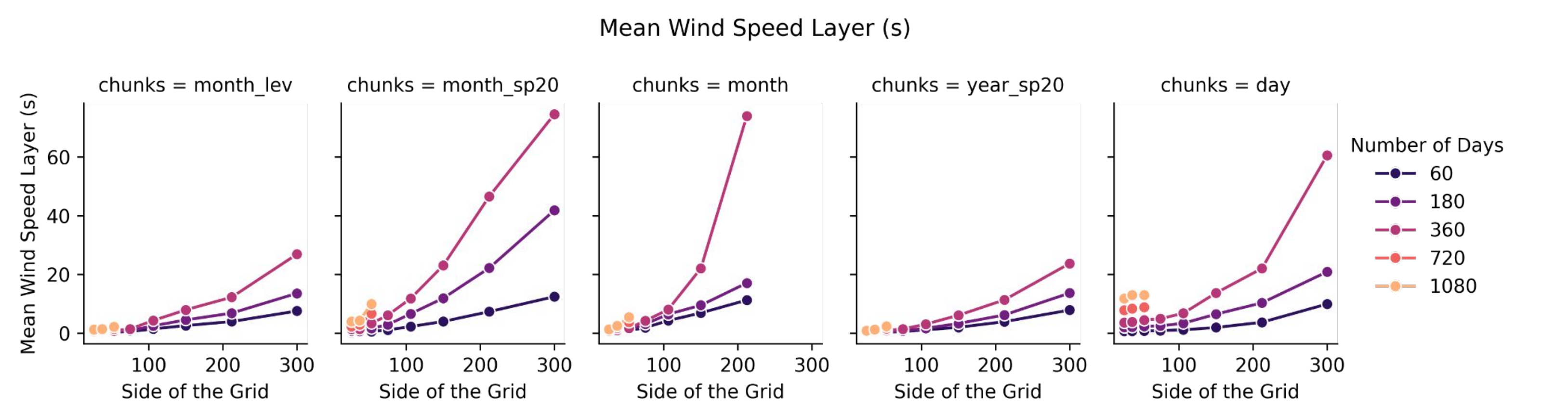




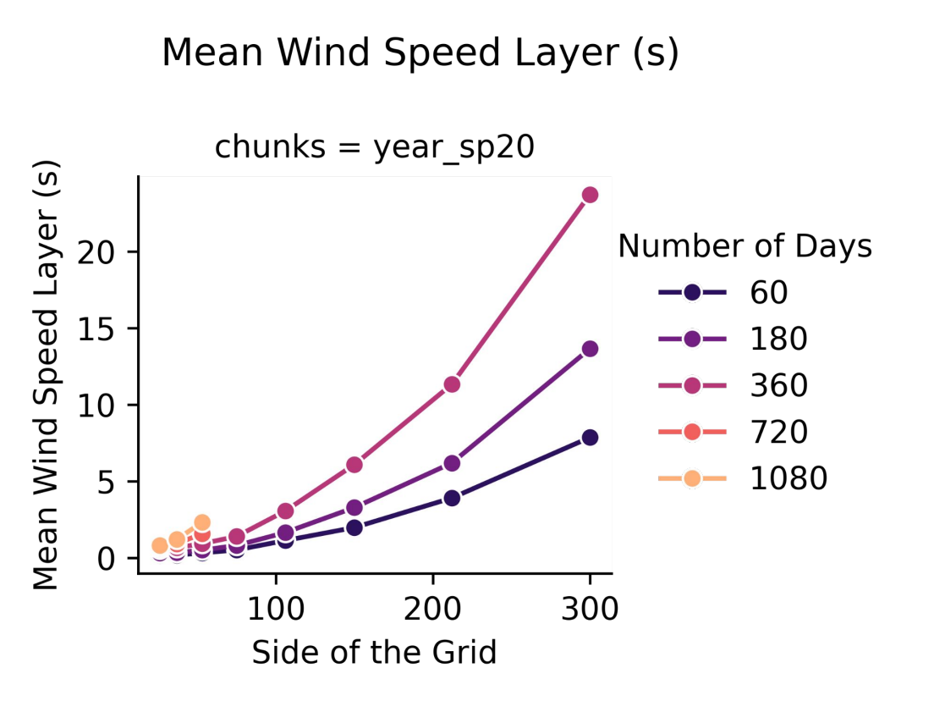
Zarr Chunking Strategies

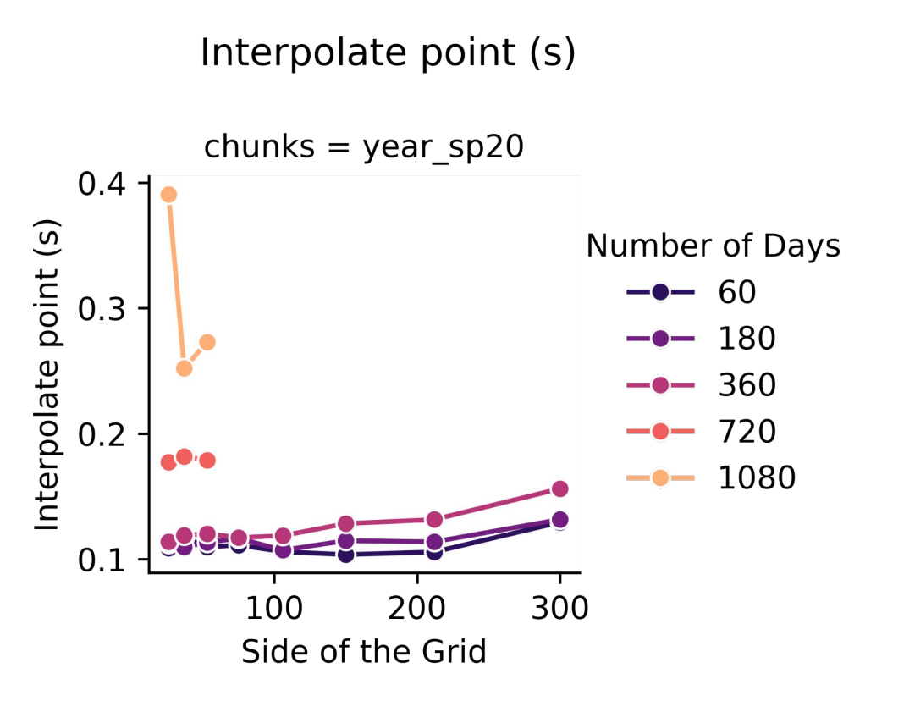
This section considers larger data sets and compares different chunking strategies for zarr files for two tasks: computing the mean wind speed over the dataset and interpolating a time series.





The 'year_sp20' strategy is the most efficient at both tasks. The interpolation of time series takes less than a second for all sizes and lengths of the data sets.





Conclusions

Zarr files, even without using a high performance parallel computation system or cloud storage, have improved every part of the process when compared with reading directly netCDF files with python. Zarr files also reduce the storage space required and, when the chunks are chosen smartly, can improve the performance several orders of magnitude.

The best chunking strategy consisted in:

- storing levels separately
- chunking space (latitude and longitude) in smaller grids
- storing longer time periods on each chunk

Summing up, the wind industry is undoubtedly moving towards wind resource assessment strategies that involve more and more data and exploring the data formats options can result in relevant improvements in performance and storage space .

1. Zarr documentation: *https://zarr.readthedocs.io/en/stable/*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 828799

Watch this presentation

Replace with QR code

Download the poster

Replace with QR code

TECHNOLOGY WORKSHOP 2021 9-10 SEPTEMBER ITALY

windeurope.org/tech2021

#WindTech21