See www.cisl.ucar.edu/events/icas-2024 for the detailed agenda and schedule.

Keynote

The technology that deletes photobombs can do climate research? The chat bot that writes poetry can do climate analysis? Future of Data Analysis

Christopher Kadow, German Climate Computing Center (DKRZ)

Climate change research today relies on climate information from the past. Historical climate records of temperature observations form global gridded datasets that are examined, for example, in IPCC reports. However, the datasets combining measurement records are sparse in the past and very low resolution. We found that recently successful image inpainting technologies, such as those found on smartphones to get rid of unwanted objects or people in photos, are useful here. The derived AI networks are able to reconstruct artificially cropped versions in the grid space for any given month using the missing values observation mask. So herewith we have found with AI a technique that gives us data from the past that we never measured with instruments. The integration of these technologies in and around climate modeling processes, particularly at the German Climate Computing Center (DKRZ), demonstrates their potential to enhance, complement, and in some cases revolutionize traditional modeling approaches. AI's role in improving the resolution of climate models through advanced downscaling techniques demonstrates its ability to refine model simulations. Deep learning techniques include U-Nets, diffusion and vision transformer models.

Climate research often requires substantial technical expertise. This involves managing data standards, various file formats, software engineering, and high-performance computing. Translating scientific questions into code that can answer them demands significant effort. The question is, why? Data analysis platforms like Freva (Kadow et al. 2021, e.g., gems.dkrz.de) aim to enhance user convenience, yet programming expertise is still required. In this context, we introduce a large language model setup and chat bot interface based on GPT-4/ChatGPT, which enables climate analysis without technical obstacles, including language barriers. Not yet, we are dealing with climate LLMs for this purpose. Dedicated natural language processing methodologies could bring this to a next level. This approach is tailored to the needs of the broader climate community, which deals with small and fast analysis to massive data sets from kilometer-scale modeling and requires a processing environment utilizing modern technologies, but addressing still society after all - such as those in the Earth Virtualization Engines (EVE - eve4climate.org).

iCAS 2024 Panels

Schedule of panels and presentations still to be determined.

International collaboration towards sustainability in weather and climate modelling

Moderator: Michele Weiland (EPCC)
Panelists: Thomas Hauser (NSF NCAR), Mark Parsons (EPCC), Tiago Quintino (ECMWF) and Ilene Carpenter (HPE)

Computational modelling and simulation of the atmosphere is a challenging and compute resource intensive domain that is vital to furthering our understanding of climate change, the factors that influence it and the impacts, including hazardous weather. Like high-performance scientific computing more broadly, it is not exempt from the need to adapt to the strategic challenges of Net Zero emissions and sustainability. Driving computational modelling and simulation, and large-scale data analysis, towards greater sustainability is crucial if the scientific community is to keep justifying the use and cost of large-scale HPC resources in the face of climate change. Electricity consumption of data centres and systems is by far the largest contributor to the carbon footprint of operational HPC; minimising the energy that is consumed and reducing/reusing waste is therefore key in achieving Net Zero.

This panel will discuss the importance of knowledge exchange through international collaboration to progress the state-of-the-art in sustainability and power/energy efficiency of computational modelling and simulation. It will bring together leading experts in data centre and systems operations, HPC experts in both hardware and software, as well as domain experts in weather and climate to debate the challenges facing the weather and climate modelling going forward. The panel will discuss how to deliver a full-stack approach to addressing the Net Zero challenge from the data centre through to individual applications using both computational and climate science expertise.

Climate modeling at Exascale: Status, Challenges and Collaboration Opportunities

Moderator: Aaron Donahue (Lawrence Livermore National Laboratory)
Panelists: Hisashi Yashiro, (National Institute for Environmental Studies, Japan), Thomas Geenen (ECMWF, Italy), Aaron Donahue (LLNL), and Anurag Dipankar (ETH Zürich)

The diversity of computer architectures and programming models pose a formidable challenge for climate model development and porting efforts. The purpose of this panel discussion is to illuminate the recent progress made in the development of climate models designed for exascale supercomputers. Furthermore, the panelists would identify common challenges encountered and explore potential avenues for collaboration.

Performance, portability and productivity for climate and weather codes in the age of accelerators: pipe dream or realistic ambition?

Moderator: John Clyne (NSF NCAR)
Panelists: Claudia Frauen (DKRZ), Aaron Donahue (LLNL), Dipankar Anurag (ETH Zurich), Segi Siso (UK Science and Technology Facilities Council)

The demand for increased realism made possible by next generation storm resolving (kilometer-scale) climate and weather models is putting tremendous pressure on model developers to port codes to GPUs and other accelerators that offer both better computational performance and energy efficiency than CPUs. Climate and weather modeling centers around the world are in various stages of adapting their codes to run on a variety of devices. Efforts to date have demonstrated encouraging improvements in model run times, but these optimizations come at significant cost due to the size and complexity of the codes involved. Furthermore, optimal performance is typically limited to a single vendor's hardware (non portable), and the maintainability of the code often suffers, reducing productivity. Proponents of language-level parallelism (e.g. do-concurrent in Fortran) and Domain Specific Languages (DSLs) claim these approaches can achieve not only acceptable performance, but portability, and developer productivity as well. However, concerns arise around the level of support for these technologies and their long term viability. This panel will explore the questions of whether achieving portability, performance and productivity is realistically possible, how it can be achieved, and how we best future-proof our investments in climate and weather model code development.

Ease of use for complex Earth System Science workflows

Moderator: Thomas Hauser (NSF NCAR)
Panelists: Rich Lawrence (UK Met Office), Tiago Quintino (ECMWF), Doug Schuster (NSF NCAR)

Challenges and Opportunities from AI workflows at HPC Centers

Moderator: Thomas Hauser (NSF NCAR)
Panelists: Ben Evans (NCI), Tsengdar Lee (NASA), Martin Palkovic (ECMWF)

---

iCAS 2024 Presentations

Diversity, Heterogeneity and Collaboration amidst Divergence

Ilene Locker Carpenter, HPE

The rapid development of AI models for NWP in the past year has brought a dramatic increase in the heterogeneity of workloads that will be used for weather and climate prediction. At the same time, collaborations often means that researchers and modeling teams need to use several quite different system architectures – different at the node level, interconnect, storage, software stack and programming environments. How will the rapid rise in AI work change HPC systems for weather and

climate centers? What software elements need to be open source and what can be proprietary? This talk will include recent examples of collaborations between HPE and its customers to develop, deploy and extend open tools like SmartSim which facilitate collaboration in a rapidly changing world.

Climate-resilient snowpack estimation with machine learning

Marianne Cowherd, University of California, Berkeley

As we progress into the 21st Century, warmer temperatures and shifting atmospheric circulation patterns are expected to lead to notable alterations in the timing, quantity, and spatial distribution of hydrological processes. This change affects water security and infrastructure and is also expected to negatively impact our ability to measure these same effects. In other words, climate change is detrimental to our ability to observe climate change. In this project, we use snowpack measurement networks in the United States as a case for exploring a) how network-based in situ measurements will change in representativity in the future and b) how we can use machine learning and gridded meteorological products to remedy the decline in sensor network usefulness. To do this, we investigate the performance of the Snow Telemetry network under a suite of dynamically downscaled CMIP6 models. We then use the WRF data to develop artificial intelligence-based models of snow distribution under future climates. Our primary focus is to discern the relative significance of model complexity and observational data in constraining snowpack estimates, especially when applied to downscaled climate models over the WUS. Our findings indicate that sufficiently intricate data-driven models are likely to sustain the accuracy of snowfall and snowpack estimates even under no-analog future climate scenarios. Additionally, these models can handle novel spatial and temporal correlations between predictors of snowfall and end-of-winter snowpack, allowing for reliable estimation in the face of unprecedented atmospheric conditions. By employing nimble artificial intelligence-based models, adept at incorporating partial and multi-modal snowpack information, we can effectively tackle these challenges. This approach ensures the resilience of end-of-winter snowpack estimation in the WUS, even as the climate evolves into uncharted territory. Lastly, we discuss strategies for similar implementations in other networks of environmental sensors; sensor networks and field measurements tend to follow political boundaries due to funding structures but hydrology does not recognize the same borders.

Towards accelerated computing in a Python framework

Anurag Dipankar, ETH Zürich

The modern exascale computing systems have given the climate and weather modeling community an opportunity for a step change in the simulation capabilities. One expects a resolution high enough to resolve the key atmospheric processes provided the machine is properly utilized. However, this is a non-trivial task. Given the complexities at various levels from the hardware to the software (model), a close collaboration between the computer scientists, software engineers, and domain experts is required to redesign the conventional model in a way that it fully utilizes the exascale capabilities while keeping the model easy to use. In EXCLAIM, we are developing a python-based framework that allows users to write code, debug, run the model, and visualize from within the framework. Computations are largely handled by the embedded DSL, GT4Py, whereas the driver code is envisioned to be in python. This approach decouples the hardware specific implementation choices made for efficiency, from the functional choices by the domain experts. The framework is expected to allow for interoperability with

software components developed in other programming languages or other frameworks. The talk will give an overview of the progress we have made in the last years and challenges we have encountered.

Efficient resource usage for large-scale earth system model simulations on heterogeneous hardware

Jan Frederik Engels and Claudia Frauen, German Climate Computing Center (DKRZ)

On current exascale HPC systems most people only focus on GPUs, but all GPU nodes also contain CPUs, which are rarely utilized. Given this and the fact that not all codes are equally suited for running on GPUs we want to discuss approaches that go beyond the strategy of just porting everything to GPUs. Modern earth system models (ESMs) are becoming more complex by integrating more components. The different components of an ESM also have different computational characteristics, with some making good use of throughput devices like GPUs, while others, for example, only solve 2D problems, which have very low computational intensity. An example of such an ESM is the ICON model, which is used for km-scale simulations on various different hardware architectures. The atmospheric component of ICON is successfully running on GPUs but other components like the ocean model have not or only partially been ported to GPUs and might also not benefit to the same degree from running on GPUs. Thus, in this proposed talk we will discuss approaches to efficiently perform km-scale earth system simulations by optimally utilizing the available hardware resources. As a computing center, we are also interested in reducing energy consumption of simulations. Beyond savings by reducing run time as described above, we will also discuss strategies to optimise energy consumption for the full machine given our standard workload. As an outlook we also briefly touch on the idea of running every ESM component on the architecture suited best to its computational profile."

Rapid emergence and applications of AI/ML at NCI in the Earth System Sciences

Ben Evans, NCI Australia

Artificial Intelligence and Machine Learning (AI/ML) has been rapidly showing its potential as a new computation approach that either replaces or augments traditional modelling and data analysis. However, despite its power, these areas are not yet fully adopted and pervasive within the earth system sciences research community or in production services. In this talk, I will describe our focus and progress in applying AI/ML, and new emerging areas that we are exploring.

Weather Prediction with ICON on GPUs

Marek Jacob, German Weather Service (DWD)

Graphics Processing Units (GPUs) have become integral components in the architecture of numerous world-leading supercomputers, offering energy efficient and powerful computational capabilities. The incorporation of GPUs, however, necessitates the adaptation of existing codes. In a collaborative international effort, the open-source weather prediction model ICON has been successfully ported to GPUs. MeteoSwiss scheduled the deployment of the ICON-GPU as its primary numerical weather prediction (NWP) system in mid-2024. Similarly, the German Weather Service (DWD) is gearing up for potential future GPU-based systems. Our presentation delves into the porting strategy of ICON using

OpenACC compiler directives, highlighting recent successful ports essential for operational production. We also share insights into performance and experiences gained from deploying ICON on GPUs, as well as on recent NEC SX-Aurora Tsubasa Vector Engines installed as part of DWD's latest expansion stage in fall 2023. In addition to the porting of the classical ICON, the DWD applies GPU technology in artificial intelligence applications. Among various specialized applications, the DWD is actively involved in enhancing its NWP capabilities through the development of the ""AICON"" system. AICON utilizes machine learning and AI technologies to fully emulate the ICON-NWP model. This presentation introduces AICON-Graph, a graph neural network implementation within the AICON framework. AICON-Graph is a collaborative effort driven forward by the DWD and its partners.

Supporting an Evolving HPC Community at NCAR

Rory Kelly, NSF NCAR

"The National Center for Atmospheric Research (NCAR) has a long and rich history with high-performance computing (HPC), having been a pivotal early adopter and having procured over 40 HPC machines since the early 1960s. Owing to this extensive history, the majority of modeling codes at NCAR have been developed for CPU architectures, although our CPU modeling community has navigated substantial technological transitions through scalar, SIMD, wide-vector, shared-memory, and distributed-memory cluster architectures. The industry's increasing adoption of GPU computing, further propelled by advancements in artificial intelligence and machine learning, is leading to another technological shift for our user community. Our latest HPC procurement, NWSC-3, was the first to mandate inclusion of a GPU-capable compute partition, accounting for a significant fraction of the machine's overall computational capacity. This presentation will cover the development of the NWSC-3 benchmark suite, aimed at procuring a machine meeting several key objectives: supporting existing CPU-based modeling workflows, encouraging our developer community to embrace code modernization for GPU architectures, and providing a substantial compute resource for our emerging GPU computing community, capable of supporting production science on the GPU partition. We will discuss the benchmarks and metrics, including the creation of the Cheyenne Sustained Equivalent Performance (CSEP) metric, used to compare the aggregate performance of a machine with both CPU and GPU-based nodes to Cheyenne, our previous CPU-based HPC cluster. We will review the procurement outcomes and look at the resultant machine, Derecho, to evaluate how well the benchmark suite achieved its goals. Additionally, we'll examine usage patterns on Derecho's CPU and GPU partitions during the first year of production, highlighting community adoption patterns, efforts to support code modernization and GPU readiness, and potential implications for a Derecho follow-on machine.

All data everywhere all at once

Richard Lawrence, The Met Office

The Met Office plans a fundamental shift in how it will curate and share data. This presentation will outline how the Met Office will change its focus to making data available followed by ensuring it is as accurate, consistent and useful as possible. Utilising the partnerships we have with the cloud hyperscalers and UK and international weather and climate agencies we look to define and share plans for increasing the reach of our weather and climate data. The presentation will close with how we see data architecture evolving in the coming years and how it will adapt to ML data driven models upending the traditional paradigms of data transfer and data proximate computation.

Challenges and Opportunities Represented by Computational Modeling and AI/ML Workloads

Tsengdar Lee, Laura Carriere, Dan Duffy, NASA

We have seen unprecedented changes in computing architecture and resources in recent years. First, managers at high-end computing centers are facing significant challenges but these also present new opportunities for scientific research. We need to support the traditional CPU based workloads, provide the resources and manpower for transitioning the legacy model codes to CPU-GPU based heterogeneous architectures, while at the same time there are more and more users demanding GPU-based systems for AI/ML workloads. These are just the technological challenges and opportunities. Programmatically, NASA Earth science has launched an Earth Science to Action (ES2A) initiative that would dramatically change the focus of our modeling program and the use of analytical tools. In this talk, we will go over the recent activities at NASA with the emphasis on the integration of modeling, on-prem computing, cloud computing, and the software stack. Join us as we present our vision toward an integrated ES2A environment.

Optimizing compute-intensive weather kernels on AMD GPUs

Paul Mullowney, AMD, Inc.

High fidelity, high accuracy weather forecasting requires the use of substantial computational resources. Many of the codes employed for this task have been developed over many years under large multi-national collaborations. There is a well justified desire to minimize changes to the underlying source code, with all its embedded knowledge, while simultaneously taking advantage of modern HPC resources. With this in mind, many of the development teams have chosen the path of directive-based offloading to leverage GPU compute architectures. In this talk, I will discuss the collaborative efforts of ECMWF and AMD to optimize the CloudSC component of the IFS forecasting model to AMD GPUs. The CloudSC compute kernels are resource (register) hungry computations that require careful implementations at all levels of the software stack. I will focus on tuning the directive-based, offloaded Fortran kernels to achieve competitive performance to native HIP implementations. This will include a detailed performance analysis using the Omniperf open source tool as well as introspection of the compiled code. We will focus on results for both MI200 and MI300 GPUs.

From Reading to Bologna, from 18-km to 9-km ensemble, from IFS to AIFS

Martin Palkovic, ECMWF

In autumn two years ago, ECMWF successfully started the operations from its new data centre in Bologna and last year before the summer successfully implemented 48r1 cycle that increased the ensemble resolution from 18 km to 9 km. This meant upgrading the 50 member ensemble to the same high-resolution as the deterministic forecast, which will be discontinued in the next cycle upgrade. After intense developments, ECMWF introduced its data driven model AIFS in autumn last year at the ECMWF HPC Workshop in Bologna and started experimental dissemination of AIFS products at the beginning of this year. The talk will cover the overview of the migration to Bologna, AIFS developments and dissemination as well as preparations for ECMWF next HPCF that should become operational at the end of 2027 and encompass hybrid IFS as well as AIFS workloads.

NVIDIA Directions in Energy Efficient HPC for Driving Earth Digital Twins

Stan Posey, NVIDIA

Efforts are underway in the weather and climate modeling community towards refining the horizontal resolution of atmosphere GCMs towards km-scale, in order to explicitly resolve certain small-scale convective cloud processes and provide more realistic local information on climate change. At the same time, Exascale HPC systems have arrived and in most cases are designed with GPU accelerator technology that offers opportunities in reasonable simulation turn-around times balanced with efficiency in energy consumption. Ultimately, output from these storm-resolving models at km-scale will become the essential driver behind the deployment of Earth digital twins for programs like the EU Destination Earth and NVIDIA Earth-2. This talk will describe advances in HPC for (i) GPU-accelerated numerical models, (ii) AI software and system features for large-scale data handling and ML model training and inference, and (iii) interactive volume visualization that together provide the critical components towards the vision of Earth system digital twins.

Pivoting to NSF NCAR's Next Generation Geoscience Data Exchange, Integrated Research Data Commons

Doug Schuster, NSF NCAR

This presentation will highlight NSF NCAR's plans to develop and deploy the Next Generation Geoscience Data Exchange, Integrated Research Data Commons (NG-GDEX). NG-GEX will provide data science infrastructure that can overcome the research challenges described above and position NSF NCAR to bolster its integration with NSF's National Discovery Cloud for Climate and be well positioned to support the National Artificial Intelligence Research Pilot Program, by connecting community model generated products with datasets produced through NSF's Facilities for Atmospheric Research and Education program. It is envisioned that NG-GDEX will democratize computational and data-driven research approaches, promote open science and broaden participation—from experienced scientists to early career researchers—through community-driven data sharing, with a focus on promoting best practices for curation of digital assets (FAIR and CARE principles, Analysis Ready (AR) and Artificial Intelligence (AI) ready data structures), and by providing simplified access to analysis-oriented resources and services through web based applications, including computational notebooks.

Future of Canada's Hydrometeorological Supercomputing Service

Charles Schwartz, Shared Services Canada

The trend in acquiring and operating larger and larger supercomputing infrastructure to support national hydrometeorological programs has been fairly consistent over the last several decades. This trend has reached its limit and is no longer sustainable given modern realities by power constraints, Moore's Law slowdown and financial considerations. Disruptive approaches have arrived that bring with them new opportunities – AI, machine learning, cloud – and new costs and considerations. We have a shared duty to help scientists model climate, while reducing the environmental impact it causes. One factor is the growing importance of data – its use in AI and machine learning is essential but storing data is costly. Migrating from one platform to another is also rapidly becoming unthinkably complex – Canada will have more than 1 Exabyte of meteorological data stored by 2025. Technologies to support AI and machine learning workloads are expensive both financially and ecologically. What does this

mean for the future of hydrometeorological supercomputing systems? Canada's current supercomputing infrastructure for hydrometeorological services is sunsetting soon and we are preparing for the future. What should the next generation systems look like? We will discuss alternatives approaches that we deemed unthinkable only a few years ago: heterogeneity of computing platforms and on-premises and cloud systems; the divergence of a forward-looking R&D environment that is not aligned with the existing operational systems to fork development of new technologies and approaches; and re-thinking our data management practices.

DART: 20 Years of Collaboration for Advancing Earth System Science

Marlee Smith, NSF NCAR

The Data Assimilation Research Testbed (DART) is an open-source software facility for ensemble data assimilation that has been at the cutting edge for over 20 years, supporting a diverse community of Earth system scientists. DA combines information from numerical forecasts with measurements of the Earth system to enhance the value of both. Applications include generating initial conditions for forecasts and predictability studies, diagnosing model error and bias, and assessing the value of existing and planned observations. This presentation will detail several impactful collaborations and further explore the capabilities and innovations in the DART software that enable, support, and welcome collaborations in the context of ever-evolving technologies, software practices, and science. Concluding discussion will focus on the importance of providing student and early-career opportunities and mentoring in research software engineering, especially due to its interdisciplinary nature at the convergence of software and science. DART interfaces with many Earth system models and observations, as diverse as ocean biogeochemistry and space weather. The software implements uniquely powerful DA algorithms like novel methods for pollutants, sea ice concentration, and soil moisture. We deploy flexible, robust tools to handle ever higher-resolution models, increasingly large numbers of ensemble members, and soaring numbers of observations. DART is also crucial to the application of innovative artificial intelligence approaches to forecasting by generating reanalysis datasets. Focusing on accessibility to support a diverse user community, our software is carefully engineered to run efficiently on systems ranging from laptops to supercomputers and be compatible with most compilers. DART facilitates adding new models and observation types and deploying state-of-the-art DA systems. We strive for the democratization of Earth system science by working with contributors ranging from high school students to renowned researchers, with institutions including the Euro-Mediterranean Centre for Climate Change, KAUST, the University of Hamburg, and the University of Vienna.

GT4Py: A Python Framework for the Development of High-Performance Weather and Climate Applications

Hannes Vogt, CSCS (ETH Zurich)

GT4Py is a Python framework for weather and climate applications simplifying the development and maintenance of high-performance codes in prototyping and production environments. GT4Py separates model development from hardware architecture dependent optimizations, instead of intermixing both together in source code, as regularly done in lower-level languages like Fortran, C, or C+. Domain scientists focus solely on numerical modeling using a declarative embedded domain specific language supporting common computational patterns of dynamical cores and physical

parametrizations. An optimizing toolchain then transforms this high-level representation into a finely-tuned implementation for the target hardware architecture. This separation of concerns allows performance engineers to implement new optimizations or support new hardware architectures without requiring changes to the application, increasing productivity for domain scientists and performance engineers alike. We will present recent developments in the project: support for interactive debugging, new compiler passes that optimize data-movement, an improved frontend with support for high-level constructs and new backends connecting GT4Py with existing HPC frameworks (DaCe, Jax).

Asynchronous IO and optimized data compression workflow

Haiying Xu, NSF NCAR

Object storage technologies are emerging because such storage offers easier access and unlimited scaling, to petabytes and beyond. These features allow object storage can meet the necessities of the large volume data of scientific computing. However, if scientists want to incorporate object store data format in their simulations, they need to put a lot of effort into modifying their simulation software. Our workflow can let simulations generate object store data or traditional file system data without significant changes. Also, using this workflow, we introduce a compression workflow to compress petabyte-scale scientific data in the fastest way by various compressors with slow or fast compression speed. We tested this workflow on the MURaM simulation model and achieved very good performance results and scalability.