CP4CS Final Workshop minutes

Attendees:

Ruth Petrie ✓
Jen Bulpett ✓
Martin Juckes ✓
Guillaume Levavasseur ✓
Matt Pryor ✓
Ag Stephens ✓
Bryan Lawrence ✓
Carsten Ebrecht ✓
Stephan Kindermann ✓
Anca Brookshaw ✓
Steger Christian ✓
Andras Horanyi ✓
Gabriellla Zsebhazi ✓
Klaus Pankatz ✓
Carlo (BSC) ✓
Antonio Cofino ✓
Francisco Doblas-Reyes ✓
Antti Makela ✓
Sylvie Joussaume ✓

Minutes

Presentations

Introduction by Ruth including laying out the agenda and introducing Martin as the meeting chair.
Ruth explained the "C3S Climate projects: and overview" slide which reviews how the different C3S projects are related, a brief description of each and their relative timings.
Sylvie asked for a few words about ESGF data services- Ruth: 34f is an extension to 34a, it is about to be submitted and is a continuation of the work that we are already doing, no new work, maintaining CMIP5 data at 3 sites and supporting CORDEX work up until the end of that project
Sylvie mentioned 34a Lot 3 in case they want to include in the slide for a more complete view
Andras- there is 34b Lot 2 which is making new regional climate runs for euro CORDEX domain, 34xx will issue an RFQ very soon
Guillaume: 34d (Regional CORDEX)- not sure if UK agrees, with data provided and relies on the same infrastructure as 34a and 34b, Ruth: important that infrastructure is maintained, Martin: we are aware of the bid but no details yet, Guillaume: bid is in final status, Andras: project has started 1st Dec, just finalising contract

WP2: Data Management (Ruth Petrie, Martin Juckes)

Ruth: Presenting slides, introduced the WP as a WP relating to data and supplying the data to CDS, aim to take a large amount of CMIP5 data available, focus on seven key experiments and 50 key variables, few extra variables for 34a_Lot2… (more details on slide 2) some data was excluded at the time from Japan, for multi-model analysis- can they use the same variable across different models -too much available data so focussed on making it all available
Slide 3- explaining the data availability matrix tool
Slide 4- Number of issues exist within the CMIP 5 archive, added quality controls checking metadata conformation and compliance with CF conventions
Slide 5- checker that no temporal information is missing (continuation of quality control), additional check for known errors in some models- these were corrected or omitted from the subset (not many)- Only data that passed all quality control was passed to CDS- only corrections made were minor metadata changes- no correction of the actual data
Slide 6: screenshot of data available in CDS- Ruth gave a quick explanation of the slide
Slide 7: Documents created about the global climate projections
Slide 8: CMIP5 continuation- have decided on one variable to definitely include, don't foresee any issues (port code and provide additional data)
Slide 9: Some predicted CMIP 6 issues- data still evolving, larger volumes so taking a copy to create a subset may not be best idea (not a stable complete set of data like CMIP5 is), missing documentation
Slide 10:
Questions: Sylvie- on CMIP 6, when a new set arriving, should be some time before they jump to have the data available, normally some time for checking, cleaning, data needs to be in good shape for several models- relatively normal, compatible with new call that will be launched, not ready for CDS necessarily, Ruth: yes, data settled down a bit by the middle of next year, Martin: we play quite a role in helping the data to settle down, we don't wait for it, Sylvie has said it is normal, it is a process that we need to plan for, make it as efficient as possible and not duplicate too much effort, quality control in a number of places (at CDS etc- need those to be aligned)
Paco- in list of variables he couldn't see ocean data- is there a plan to include this in the next contract? Ruth: no ocean data was supplied on the ocean grid, sea ice thickness etc was provided as it was produced on the atmospheric grid, no plans to include anything other than that at this stage, there is a page for the documentation for the data repository, Paco: I would like to know what the process is for moving the documentation from ES-Doc to the model documentation, Ruth: no dynamic link, the model info that was put into the Copernicus web page was a manual process, would be good if it could be automated. No formal link between ES doc and the documentation at Copernicus Martin: ESdoc has gone through a significant upgrade in the last few years- should be easier to do something systematic/automated for CMIP6, Ruth: can revisit it and how it is linked into the CDS, Paco: asking as they are responsible for the QC of the datasets in the datastore- there is a documentation aspect, want it to be compatible with CDS etc and making it available, may be some info that is not available in the documentation that we provide that they may need to pick up from ES doc- assumer we would want to know this process, especially for CMIP6, Martin: waiting to see the RFQ, believe it to be a short contract, greater opportunity for adjusting and lining these up will be the renewal going into the next phase of Copernicus, will be in contact in the future about QC, Byan: fixing the doc available as not always good, Paco: recommend to go back to modellers? Bryan: keep ESdoc team in the loop with issues
Martin: There is time planned in for discussion and general questions at the end

WP3: Data Node Software (Guillaume Levavasseur)

Slide 2: status of the infrastructure, describing the info on the slide, load balancing to ensure requirements for node uptime are met, data published at three sites needed development on the ESGF stack, need to ensure the same indexes are exposed for CDS
Slide 3: issues and constraints for the future- manual/bash scripts used usually but they have been using Ansible without impacts on the load balancing infrastructure, and believe that Ansible is better than Bash going forwards.
Slide 4 and 5: containerisation of the ESGF stack for scalability- important for more and more subsets for the CDS, need more resilient infrastructure and node deployments- relies on different docker images to containerise different ESGF components, microservices to control the stack, and they have successfully published this approach, will be moving from current RPM-based deployment to Docker next year, can use Kubernetes or Ansible as previously mentioned, Kubernetes used for failover could use for CDS nodes
Slide 6: some difficulties with configurations coded into components, major constraint is resource change, so they need Matt Pryor's time for now
Slide 7: highlighted the risk of Python 2 retirement at end of 2019, security updates on the OS used in ESGF, updates being carried out and IPSL have created a procedure for CentOS update
Questions: Martin: mention that some of the work to move to Python 3, hoping to cover with continuation as no funding in earlier projects to cover this.

WP4: Compute Node Software (Stephan Kindermann, Carsten Ehbrecht)

Carsten: talking through slide 2, they used MAGIC toolbox but next will want subsetting, regridding procedure
Slide 3: what they have done so far- created a cookiecutter template to generate the PyWPS…
Slide 4: for production it needs more components, more steps to use so Ansible used for deployment, can customise this
Slide 5: If there are WPS servers then can execute the drop requests on the server itself, they have created a backend to allow delegation to a scheduler for scalability
Slide 6: Birdy- not necessary to run the service but makes it more easy to work with the servers- point to endpoint and call WPS process like normal Python functions
Slide 7- WPS has no security protocol, so have an additional proxy for this
Slide 8: improvements- all works and is used, but need to update templates and playbooks, they have prepared the docker deployment but it has not been rolled out to scale with Kubernetes, may be special requests for CDS for this library, want to use Keycloak and currently trying to connect these things
Slide 9: main component to rely on is PyWPS protocol- bugs to fix, need to make a number of improvements, need to switch to a different scheduler as not well maintained, would like a new WPS interface
Step 10: data is heterogeneous so need further work to make it reliable
Questions: Martin: worth commenting that this is very important for the quality of the service that we deliver through the CDS, at the moment the data that Ruth uses is available through CDS but not the CDS toolbox- WPS layer is an essential foundation for the work that we hope to do next year to resolve this

Paco: this is the engine that is behind the provision of the projections, do we have any work to assess the user experience e.g. speed at which files are served? Any transformations of the files served compared to the ones hosted by ESGF- do you look at what the user sees at the other site? Carsten: no feedback from CDS about how they use the interfaces, Martin: no feedback, project was designed to support MAGIC and we have feedback from them, but it didn't develop into a user-facing service, in other projects there is a lot more information about the components, and the components have been more rigorously tested, Paco: asking if we ever go into the CDS to make requests ourselves to have an impression of what the user actually sees from our work, as the user won't see your work only be aware whether CDS serves the files quickly enough, QC for the filenames- does this mean that the user gets files back which are CMIP5 compliant with their names? Martin: don't think users can see any of this subsetting or regridding, it is implemented in the services and we need to build the interface to the climate store, currently users see the files as Ruth has corrected them, do not currently monitor the CDS, something to be in contact with Paco with in the future, Andras: some metrics were created for the project, there was one deliverable about how the data is used by the users, Martin: Matt may say more about diagnostics on the users, but this is a work in progress.

WP5: Tool and Services (Ag Stephens)

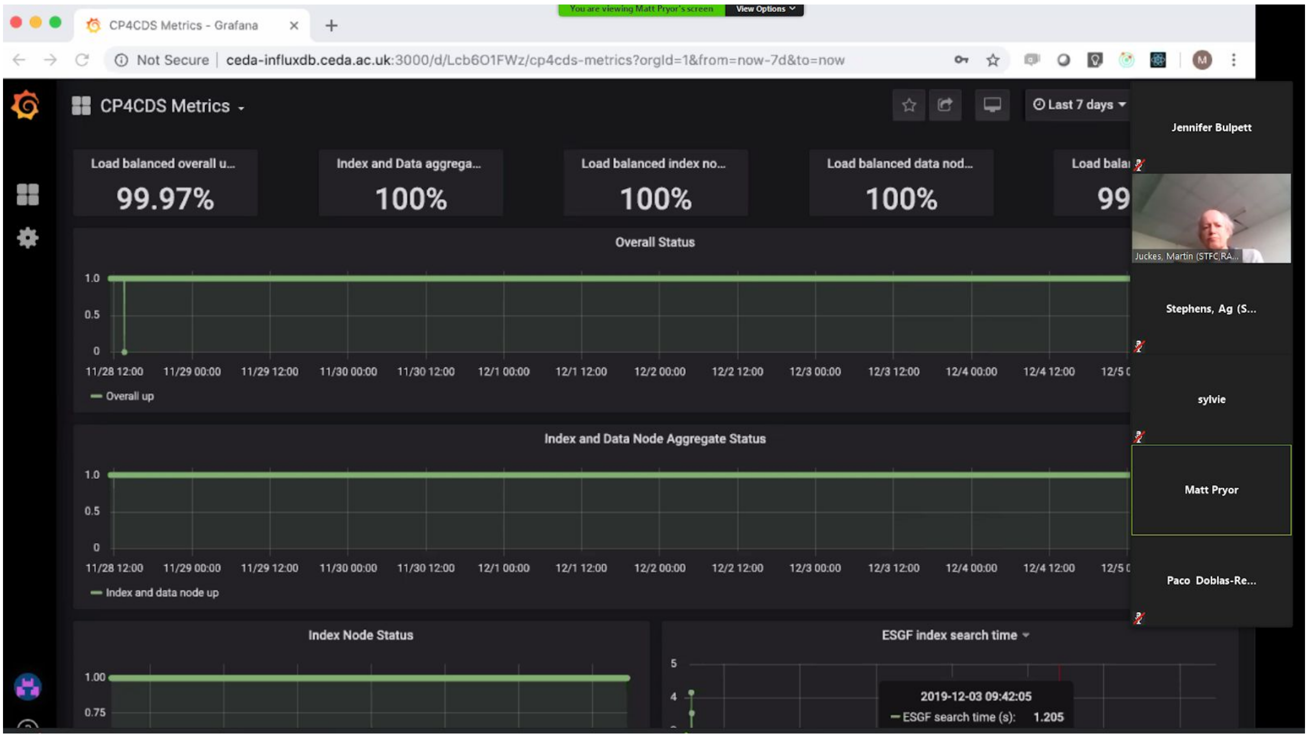
Ag introduced himself,
Slide 2: 5 key components/tasks within this WP
Slide 3: ESGF - PyClient queries directly the ESGF search API
Slide 4: During the project continued to update the codebase, updated test suite for Pytest, queried CMIP5 holdings, Jupyter notebooks to provide an example of usage for the tool, used in the publication of CP4CDS data sets, made sure it will work with CMIP6, used widely
Slide 5: Synda is developed by IPSL, use command line to define download/replication, can use with post-processing workflows, works with many protocols useful for inter-continental replications
Slide 6: Continued use of Synda including within ESGF for all major data sets, have maintained and continued to fix bugs, improved integration of post-processing, can use with Conda, supports all major datasets.
Slide 7: Task 5.3, making software environments shareable and reusable, built on Conda as its a common packaging tool within the community, conda-forge to ensure that they conform to a given dependency tree.
Slide 8: Task 5.3 Closely working with WP4, important as gives a way to deploy a fixed environment consistently on different nodes and different sites, control and support reproducibility, can say what software deployment looks like on a given.
Slide 9: Visual view of what they are trying to achieve with SddS and compute node, demonstrating how can go from an initial template (CP4CDS WPS)- developer can take a fork of a template, a software environment, modify to build their own WPS application, then they define their own environments, make changes in their own repository and then deploy into our system (includes review and testing), our system can then take a fork of their own repository and feedback into the compute structure- idea to streamline how to bring new code into the system by capturing the environment and building things on top of templates so easier to work with them.
Slide 10: Task 5.4- initially defined as extending Data Ref Syntax, as ideas developed realised what they needed was a way for the client who works with the WPS to be able to resolve inputs, need a way to work with dynamic inputs to a processing service, may be impact on selections already made, certain number of variables available based on a given selection- can work interactively with the service.

Slide 11: Came up with concept of meta-WPS to complement the main WPS, includes tags to point to meta-WPS, key advantage is that when the client is working out what to do it can continue to send request to the meta-WPS and it gives exact state at this point in time, approach that trying to reuse WPS standard.
Slide 12: Task 5.5, about proving that it was possible to use the SDDS to deploy, focussed on CliMAF from IPSL- large framework with a number of different tools to analyse and process climate data, interfaces to various different ESGF data sets, generating indices.
Slide 14- to prove end to end capability, used a subset of CliMAF, selected CMIP5 model and final output is a PNG file line graph, creates a slice of a use case.
Slide 15: Demo is available, able to deploy at DKRZ and CEDA and demonstrate end to end process.
Slide 16: To conclude, important that there was a component of CpCDS to support tools, good to extend and support those products, good foundation for 34e and the work on declaring data requirements in code has progressed understanding and will impact API design for 34e, interactions between Lot 2 and CliMAF have shown where work needed
Questions: Sylvie: Where you mention development of a common and robust solutions- will this be addressed in 34e, what are the limitations, Ag: think we can take all of our learning and all of our infrastructure into 34e, focussing on robust and resilient, basic WPS processes, subsetting, averaging nd some regridding, based on the whole birdhouse WPS framework and Ansible deployments in place have the availability to deploy next to the data, next focus is the WPS processes themselves, identified it is easy to put up a process that works on a small amount of data/few models, requirement we perceive is that it is tested with a large coverage of those projects e.g. temporal and spatial subsetting, ideally would work on 100% of the data- can't test that, investigating with 34e on a test suite to sample data sets and find heterogeneity and we can analyse those and ensure code works with them to get round data quirks, Martin: important gap atm is that this work doesn't link to CDS toolbox, in 34e this gap should be closed so that services can be accessed through the toolbox, Ag: "have we viewed datasets through the eyes of the C3S users"- work we plan on doing here, potentially means that C3S users may be requesting whole files/subsets, somewhere we need to work out how to funnel the requests to data nodes or WPS, user would just want to select the output, need a library in CDS to make these decisions

WP6: System Integration and Operations (Matt Pryor, Philip Kershaw)

Matt:
Slide 2:Initial issue to address was how to deliver CMIP5 data to C3S with >=98% uptime, developed. 3 geographically distributed sites with DNS load balancing (AWS Route 53) between them, when they make the data request they get a different IP each time, set with a low time to live, each site has a health check in route 53, allows to be removed from the pool of IPs if the health check returns as false, Matt is able to tell people when services are down but this hasn't affected service as they have been removed from the load balancing pool.
Slide 3: Resiliences, synchronisation, could reuse technology, the publication does not need to be highly available but replication does, simple using different parts of the ESGF stack, synchronisation of the data itself more difficult, replicating still requires close coordination between sites, could lower this with public cloud, cost relatively high compared to hosting onsite
Slide 4: Going forwards to deliver similar service fr CMIP6- think the same solution can be used, already being used for CORDEX, may be easier as no access control so no need to register, meeting recently with Guillaume etc- architecture review- most pressing is that Docker will be primary delivery mechanism either using Ansible or Kubernetes and potentially Autoscaling,

probably big sites using Kubernetes, small using Docker, should bring resilience with scaling, and increase portability into the public cloud, desire to use other search backends other than Apache SOLR, looking at Open search and Intake etc. In terms of data delivery- may move away from threads and provide a URI for a piece of data, would lose aggregation capabilities- still tbc, no concrete plans except for docker, Metrics- set were agreed and have been gathering since October, Matt shared a live view of the metrics dashboard



Shows last 7 days of uptime, and further down shows Index and compute node, the average time for representative search, OpenDAP and WPS request, can aggregate metrics over different time periods

Questions: Paco: does the CDS team have access to the dashboard, Matt: No, going specifically by the milestone we only need to provide the metrics for monthly reports not the actual dashboard, we could potentially share this with a small amount of work

General Questions: Martin: wanted to comment on Paco's questions on evaluations and quality control, started supporting a lot of work in MAGIC calculating indices- these then didn't match the expectations of the users of the CDS, these are central to the quality control of the climate projections- question about where to go in the future as a complete view of the quality of the climate data will need to see how the MAGIC work fits into the C3S operation as a whole

Antti: comment on MAGIC, would like to dig a bit deeper in their project about what the users want in terms of metrics, they provided one report, quite a practical description of tools, ways how the users would like to see how different models behave in respect to each other

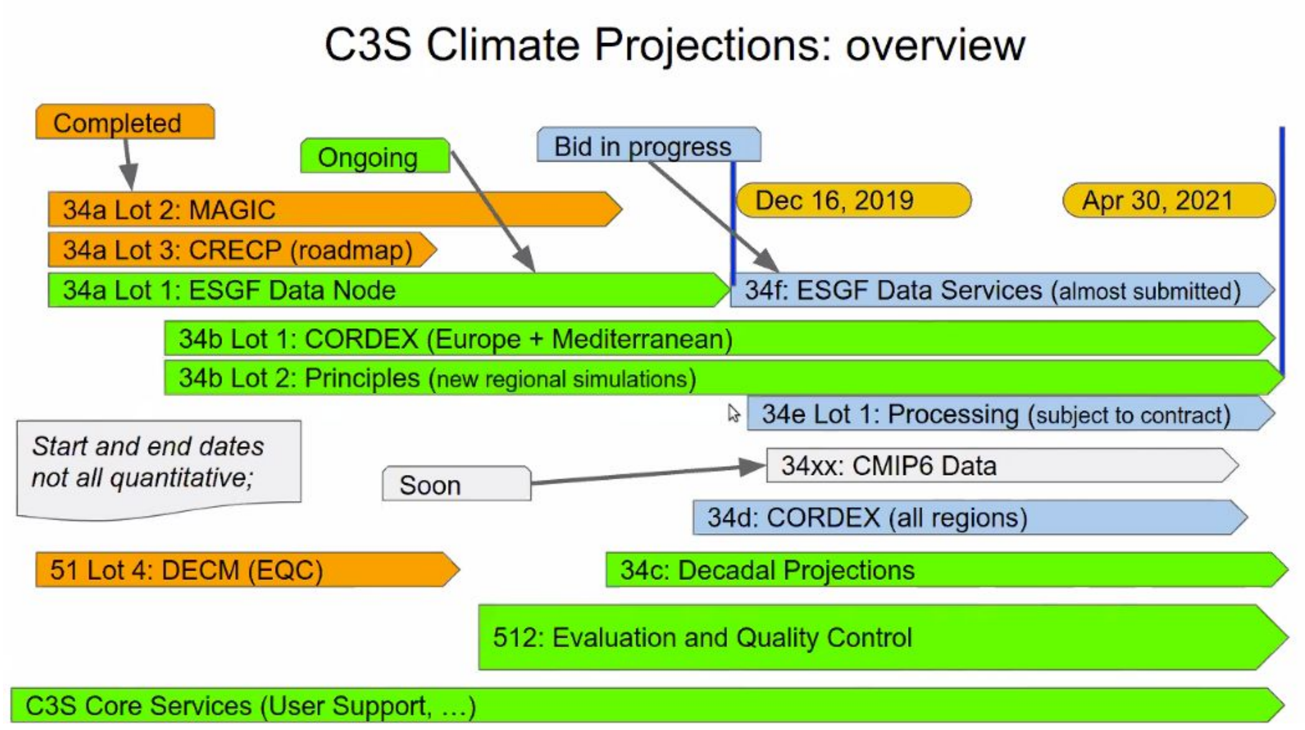
Christian: works with Paco- question to overall workflow and quality control, lots of checks made on the data before implemented into the datastore, now we have to make all these things again as not sure what happens in the CDS itself- there might be changes in the data that the user gets from CDS, for CMIP6 should think about this and not do the same things twice, either all datasets should be checked before they enter CDS, or data should be checked when user retrieves it from CDS, maybe not all checks needed before and after, need to coordinate with C3S to find out what happens in between and what to re-check, Martin: users get files that have been checked and

unless a serious fault in the system they should not be corrupted, will change when we offer data processing and will make it more difficult
Martin: do plan to explicitly provide resources to interact with the EQC function in future project plans, Christain: good to get in contact before that work starts to coordinate work, Martin: cannot talk before due to short timescales but right at the beginning.

Directed discussion/feedback

Lessons learned (project interactions, technology choices, requirements)

Martin has updated the overview of Climate Projections slide to show the additional project as mentioned earlier



C3S Climate Projections: overview

Completed

Ongoing

Bid in progress

Dec 16, 2019

Apr 30, 2021

34a Lot 2: MAGIC

34a Lot 3: CRECP (roadmap)

34a Lot 1: ESGF Data Node

34f: ESGF Data Services (almost submitted)

34b Lot 1: CORDEX (Europe + Mediterranean)

34b Lot 2: Principles (new regional simulations)

34e Lot 1: Processing (subject to contract)

Start and end dates not all quantitative;

Soon

34xx: CMIP6 Data

34d: CORDEX (all regions)

51 Lot 4: DECM (EQC)

34c: Decadal Projections

512: Evaluation and Quality Control

C3S Core Services (User Support, …)

Antti: how has the project contributed to the CMIP6 development, e.g. they have sent some examples of special cases where there were problems in the data, how has CMIP6 learned from this project? Martin: been a lot of transfer on the technology side, Ag, Guillaume and Matt could say more on making the ESGF software more resilient and responsive to improve dissemination, on the data requirements side… Ruth's quality control work has fed back, Ruth: not directly impacted CMIP6 in that sense but has highlighted problems in CMIP5 and that these shouldn't be repeated in CMIP6 but has not prevented issues with data in CMIP6, Martin: we have been doing quality control on a fairly static set of data, problem with CMIP6 is that it is a dynamic set of data, Ruth: and the timing factor as well, Martin: did mention to Sylvie earlier that hoping next year can start doing a bit more work aligning the quality control work for C3S with the CMIP6 publications workflow and that will improve the synergies and the feedback between work funded by C3S and outside c3S, Sylvie: when we have found QC issues- have these been documented and put somewhere in CMIP5 and ESGF so that we keep the information? Ruth: Has a list of log files where things have failed for specific reasons, would be a big effort to link that back to ESGF, Martin: quite a big discussion on the ECMWF infrastructure panel- data isn't their data it belongs to the modelling centers, system they had in place for CMIP6 is that people send issues to the modelling centres and they can then register/record any issues in a publicly viewable place, had something similar for CMIP5 but doesn't work well as the issues can disappear into the internal maze of the modelling

system, working on a system for CMIP6 that makes issues visible to users, Guillaume: agrees that the service in place for CMIP6 is there, but for CMIP5 but don’t have all the same information, mainly relies on climate modelling groups efforts- they are more focussed on CMIP6 current;y not CMIP5, would be useful to have the issues as modelling groups are not ready to do this. Sylvie: things that Ruth has identified- have you sent these to the modelling groups Ruth: yes but had no response as they are busy with CMIP6, no scope to go back in time that far, Martin: not that modelling centers are not interested in QC, they have extensive procedures and have responded to problems in CMIP5 before CMIP6 kicked off, we have no visibility of these as in their databases, their QC systems not designed to work with others.
Antti: for example, if his group found peculiarities they have had rapid response from the modelling centers, with C3S what is the procedure if using CMIP6 in the future- something has passed all QC checks- how do we then inform of issues, very relevant that C3S data should be well quality checked, but always remains some strange things in the data, Anca: User support is the keyword, sounds easier than it is meant to be, mechanism for collecting this information is user support, for projects currently running this has to be done, report to user support, it gets collected, labelled and tagged in a particular way, they decide with EQC how it is presented to users and how to act upon it if it is acted on at all
Guillaume: take care not to multiply the blocking issues for the whole of CMIP6 errata reporting services, they have something for ESGF if they need to do this for C3S there will be duplication, can they merge this if data served by 2 different portals, Anca: to be decided by ESGF and C3S, needs to be straightforward to the user, channel between them and pass data between in the right format, cannot report data there, Matt: Make sure that issues go into a global system not just the C3S user support system, Anca: just creates a repository, not a global solution, how do we record the CMIP5 errors currently, Martin: level of support for the community for this data is quite limited, integrated service for CMIP6 is not possible for CMIP5 as underlying technology is just not there, C3S needs to take responsibility for data for their uses, Anca: an issue that may not be resolved but we know it is there, Martin: don’t have the capacity to upgrade CMIP5 files to make compatible with CMIP6 procedures, talking about learning lessons and doing better for CMIP6 and other future projects, Ruth: as long as any issues reported to C3S are then fed back to Errata service then that should be sufficient, Matt: provides an API that can be consumed … Martin: users can’t see where an error is coming from, up to C3S, Matt: user interface could show where errors are coming through, Guillaume: talked about having an API to provide annotations about issues for each dataset, with a huge number of datasets a request has to go to the API for each dataset, could be a lot, Martin: Ag how do you feel about building this into the WPS, Ag: One of the things that we are doing in 34e is the characterisation register- mainly focussed on metadata and structural issues, not scientific ones- have a way of collecting and flagging into a public repository
Martin: find out what Paco thinks about making use of information that is in Guillaume’s Errata registry, Anca: Could ask this to Paco (and the EQC project) as a whole- put on their agenda as a whole, Guillaume: also a topic for the CORDEX projects 34b for the next year, may be suitable for CMIP5 and CMIP6, Anca: 34b Lot 1 should make recommendations for this, but down to others to make final set of recommendations, makes it very important that 34b states these recommendations clearly to lead down the path to the correct decision.
Martin: work to do around exploiting the information around errors in the data
Sylvie: mentioned that the interface with the CDS toolbox is not functional, what is the link to other aspects of CDS incl MAGIC, Martin: CDS toolbox mainly works with local data, interim library that allows people to access data in the notebook database so that they can do processing there, at the moment that doesn’t link through to ESGF data services so people have to download data to their local server and then upload to ECMWF, not a very efficient process, 34e will make that possible by

providing an interface that will couple with the toolkit, make integration smoother between the different components
Sylvie: what about the link with MAGIC, working together, Anca: we have not integrated with CDS and have no plan to, Martin: run at DKRZ, key part of evaluation of CMIP6 leading into the assessment, not the MAGIC system as a whole but that produces the metrics that were embedded in the MAGIC software., what is continuing is the environment that allowed MAGIC to run across the three data nodes will be used to run the cut-down data service that Ag is creating for 34e, the metrics that MAGIC produces are a very important part of the valuation of climate data, talked about QC but not about evaluating the climate data e.g. relative strengths and weaknesses of the models that go into an ensemble- MAGIC would produce these, uncertainty how to feed into C3S, Anca: do you know that this will use the same as MAGIC, last heard from Paco that data is so heterogeneous that metrics would not work, Anka can link with QC contacts to metrics, Christian: will use some but not all, there will be a scientific aspect, Martin: so a component of the metrics that MAGIC used will be used for 512

Opportunities (potential applications, additional data, project synergies)

Martin: planning to produce CMIP6 data into the CDS, recognising the issues that others have mentioned already data variable, data collection changing quite quickly, the question of whether there are any other data sets that we should be thinking about and the synergies with other projects, how should we communicate with the EQC work, are their others that people would like to raise?
Sylvie: possible implications of evolution of the architecture and how this impacts other projects, was mentioned that some of the software stack is still in Python 2, Martin: still an issue to address, not supported after 2019, software will still work but increasingly difficult to maintain a secure service as not necessarily able to upgrade things, institutions are working to address this within projects, hopefully will be resolved reasonably soon, other issues around having multiple copies of the data due to the fact that the quality control requirements for C3S are higher than for the global ESGF, need to talk within IS-ENES about reducing the level of duplication either by improving quality on data nodes or by making things more flexible- both of these should be raised in the architecture document, Issues around duplication of services and providing data to different people, avoiding duplication when different requirements and constraints from C3S and global ESGF, Sylvie: Question not on replication, more on evolution in the architecture on C3S, wondering on which timescale and how it might affect C3S, more on the ESGF software, Martin: we do have duplication in the software stack, there are differences, Matt: key is making sure that everyone can easily deploy the software and making upgrades easy, hopefully work being done around the docker deployment should make that easier, historically been tricky to update ESGF nodes, Martin: at the software level an increasing number of components are reusable so situation there is improving, working to make software more robust and with some degree of convergence, in terms of WPS, hoping things will go in the other direction- no focussed funding for ESGF WPS processes, work being done for C3S may be a leading candidate for ESGF WPS, need to get it working for C3S first.
Ruth: instead of the process used at the moment, could share updated lists of datasets amongst the load balancing sides to avoid too much duplication, just one or two files changed rather than large amounts of duplication, load balancing system pointing to new ones as well, not sure exactly how it would work, would have a copy of the data at every node, but instead of data duplication as present, each note would republish with a new C3S number and replicate that way then send to CDS, Matt: makes a lot of sense, reusing ESGF infrastructure rather than running our own, Ruth: not sure how that affects, would be pointing to our data storage? Matt: wouldn't have exact control of timing so would have less control of load balancing, depends on how much consistency we need, Martin: firm

requirement from c3s about knowing what is going out at each moment in time, Matt: already accepting an eventual level of consistency in that anything searchable is downloadable, may be some things that haven't been indexed, everything available in the search is available at every site, Ruth: don't think there is a latency issue, file provided to say which files are new, wouldn't do that until all 3 sites have replicated and republished to the main ESGF network, would point to old data then update manifest when all sites updated, Martin: current situation is that we have duplicate copy at CEDA and then files presented for C3S- 95% the same 5% changed, would be a big saving if only had to store the 5%, as well as saving duplicating around Europe, certainly things could do there that rely on streamlining the ESGF data management processes, issue raised earlier about getting data fixed- CMIP5- Ruth has fixed metadata errors as the modelling centres too overloaded with CMIP6 to fix errors on CMIP5, hope that we can pass errors back to the modelling centres with CMIP6 to get fixed, hopefully be in process of that data publication even if a small delay so the data is cleaner overall. Means the work that Guillaume is doing around the registration service for CMIP6 will be very important, Guillaume: make it available for users directly instead of being private for the modelling groups, open to users and then have background validation for the climate groups, could be used with CDS data, Martin: possible for others using CMIP6 data to send to Errata? Guillaume: may make the API open, but may leave the groups open to many issues, goal to make the system open to users- high requirement for CMIP6 users, not just CDS, need to develop in the next month or so. Martin: Big opportunity coming up

Sylvie: this project is extending what can be done with WPS and other facilities, what is the long term view, not clear for her, Ag: one of the requirements for the CP4CDS project was to develop a service where we could provide uptime that was higher than any of the individual platforms, part of this is developing the geographical load balancing, one site can go down the others stand up, useful approach to take forwards for general future architecture, looking to build these services- processing done on local data, e.g. beside an ESGF data node, model looking to use, can deploy for existing C3S data nodes, compute node will sit next to and have access to local file system, could equally be deployed next to the ESGF entire holdings of CMP6, because of work using Ansible and providing recipes to deploy the services, keen on idea that it could become a generic ESGF service, we have a subset of processing, can provide another ESGF partner with access to the recipe, provide e.g. 3 VMs and it will deploy everything they need to have a WPS service running with a scheduler that will then run on the batch nodes, vision is to provide a standardised approach to be used by anyone in ESGF. Sylvie: may be demanding in terms of resources, is there a business plan to sustain it? Ag: initially agrees that there is a resource hit, ongoing issue with maintenance and testing against the datasets, actual resource issue for processing and bandwidth, would argue that currently all scientists are downloading files and this is using a large amount of space, would be providing data reduction services and should free up other parts of the system, Martin: part of the design that has gone into this and processing has a relatively low maintenance cost but a high impact on users and improvements, Ag: we expect to need to keep finding resource to support and improve this- common problem with ESGF, Martin: final point is that it is essential as we cannot deliver a meaningful service to users in the C3S environment without this so we have to do it. Sylvie: if it requires a lot of computing and local storage so hardware that may cause some issues, already have 20000 users and then Copernicus could greatly increase user numbers, Martin: efficiency talks about hardware as well, downloading uses a lot of hardware as well, complex balance between them, cannot give any guarantees as don't know the load, Ag: Can manage the load as all of these processes are asynchronous in nature, they can poll the server to see how things are progressing, we can limit the number of jobs running and with the recipes can scale horizontally, adding more VMs to a cluster compared to the cost of human intervention is still a smaller cost than adding more resource to it.

Martin: lots of these discussions will be ongoing and the conversation will continue in other meetings, been a very useful discussion and helps in deciding our next steps. Any last closing remarks?
Ruth: Thanks very much everyone.
Sylvie: thanks for organising the workshop, important that Copernicus can link and important work that you have done, I want to thank the team.
Anca: With the continuation work we have in place we will try to complete the connection and see how to approach the problem in Copernicus too, we have lots of points of contact to refine this better in the next few months.
Meeting finished at 1.05 pm