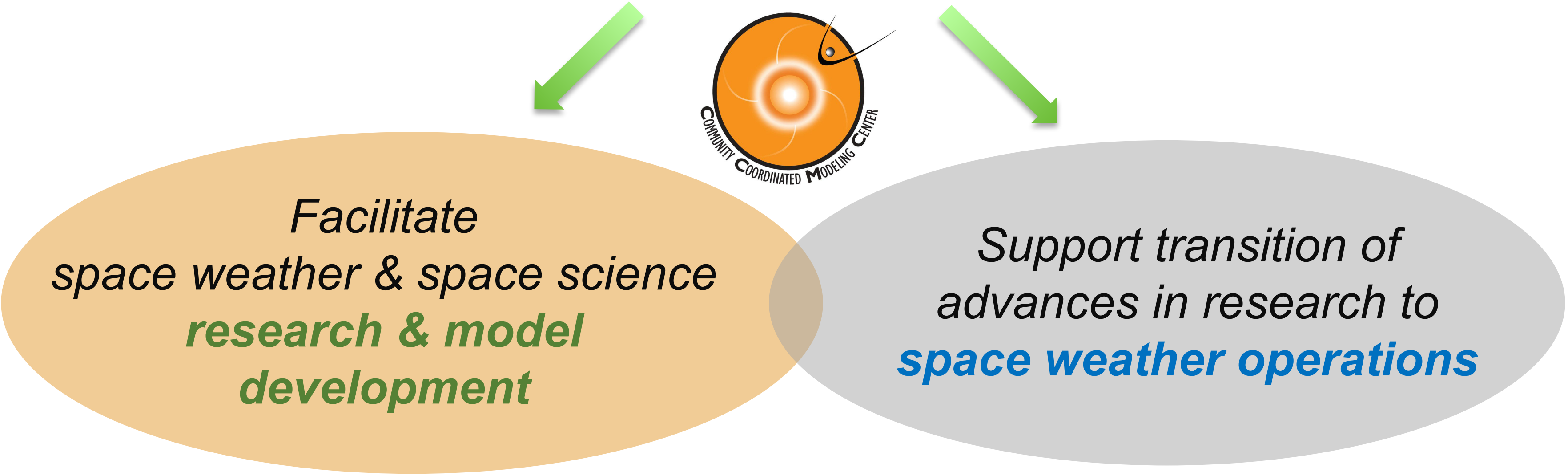
Streamlining Data Management and Operations with Airflow for Heliophysics Research and Space Weather Forecasting

Presented by Damian Barrous-Dume

CCMC/Navteca LLC

Multi-agency strategic investment in US space weather program

CCMC Goals



Facilitate space weather & space science ***research & model development***

Support transition of advances in research to ***space weather operations***

Established in 2000

Original team: Michael Hesse (founding director), Masha Kuznetsova (deputy), Lutz Rastaetter

First equipment: 3 Sun Workstations
First model: SWMF (U. Michigan)

Models at CCMC

SWMF.SC+EEGGL+CME

AWSoM EEGGL SRPM

PFSS.Petrie ANMHD

PFSS.Macneice FLAMPA

PFSS.Luhman
n

MagPy

MAG4 UMASEP SPRINTS-SEP

ASAP ASSA AMOS

WSA NLFFF GSU All Clear

MAGIC SNB3GEO FISM2

GCR BON NOVICE

NAIRAS CARI-7

Corona

WSA-ENLIL

WSA-ENLIL+Cone

WSA-ENLIL+EPREM

WSA-ENLIL+SEPMOD

HESPERIA REleASE

PREDICCS EMMREM

SEPSTER

iPATH ZEUS+IPATH

SAWS-ASPECS

CORHEL

CORHEL- CME

Heltomo IPS

GAMERA/Helio

DBM SEPSTER2D

SWMF.SH

DIPS

Heliosphere

MAGE/GAMERA+REMIX+RCM

LFM-MIX GIC

OpenGGCM+CTIM

SWMF+RCM+deltaB

SWMF+RCM

SWMF+RCM+RBE

SWMF+RCM+CRCM

LFM-MIX-TIEGCM

WINDMI LANLstar

IGRF Tsyganenko

PS VP Weigel-deltaB

AACGM Apex

AMPS GUMICS

Magnetosphere

VPIC

PAMHD

PIC-Hesse

Local Physics

RCM

VERB

AMPS

Fok.CIMI

Li's Rad Belt

PINE BSPM

UPOS RB

AE-8/AP-8

AE-9/AP-9

IMPTAM

RAM-SCB

SHELLS

ORIENT

Inner
Magnetosphere

WACCM-X WAM-IPE

SAMI3/WACCM-X

NCAR DART

GMAT CTIPe

DTM2020 IDA4D

TIE-GCM USU-GAIM

SAM SWACI-TEC

ABBYNormal

NRLMSISE

SAMI-3 GITM

PBMOD

WBMOD

Weimer IE

Weimer-deltaB

IRI JB2008

IMPACT

COSGROVE-PF

Ovation Prime

TRIPL-DA

Ionosphere/
Thermosphere

Services at the CCMC

We offer a range of run services that contribute to building a network of simulation services and run result archives. Our services support forecasting, model validation, education, and space science research.

We provide multiple different ways to perform model runs:

Runs-on-Request

Long running simulations by request. Submit a request and get notified when the results are ready.

Continuous Run

Continuously running models. View the latest results of several models without a separate request.

Instant Run

Instantly executing models. Select a model and see the results instantly in your browser.

CCMC's Hybrid Environment

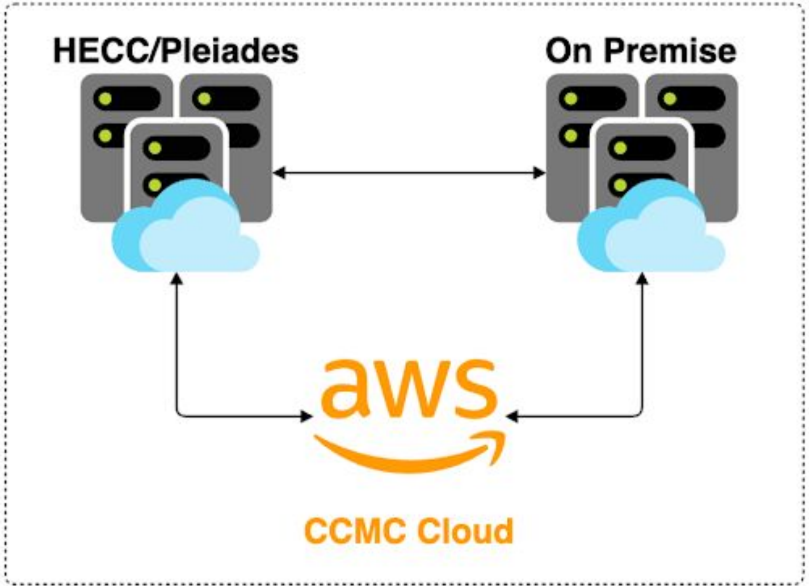
We utilize a Hybrid architecture to help manage all of the different services we help support.

But how do we manage all of the data?

Taylor-SQL
@tayloramurphy

you need DataOps so you don't make a DataOops

12:46 PM · Dec 19, 2021



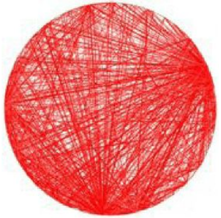
Data and Operational Challenges

Current Challenges

- Manage multiple file transfer pipelines and syncs across environments.
- Improve visibility for scattered cron jobs to ensure proper execution.
- Increase overall reliability for users and data archives.

What's the Impact?

- Inefficient operations and delays in delivering critical data to the services that might need them.



Big Ball of Mud of Pipelines

Look Familiar?

Future Challenges: The Reality of Big Data

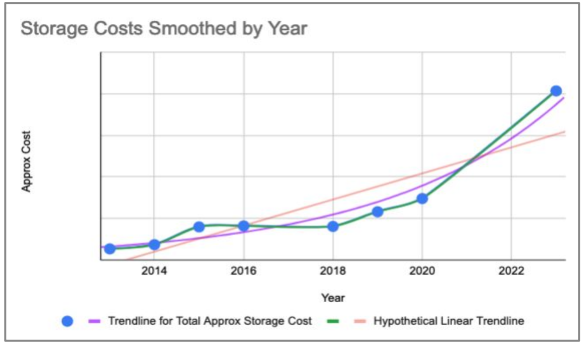
Big Data

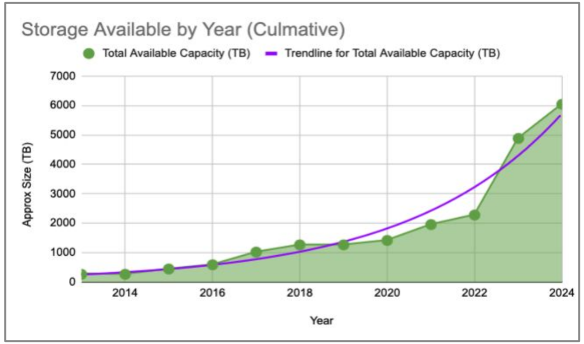
Volume
The large amount of data generated and collected.

Velocity
The speed at which data is generated, collected, and processed.

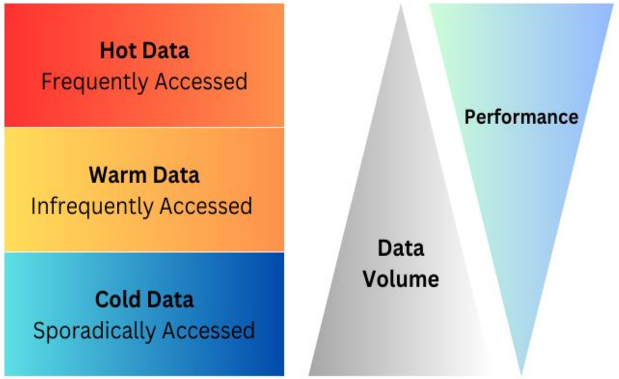
Variety
The different types of data (structured, unstructured, semi-structured).

More Data, More Costs


Storage Costs Smoothed by Year


Storage Available by Year (Culmative)

Data Tiering



Hot Data
Frequently Accessed

Warm Data
Infrequently Accessed

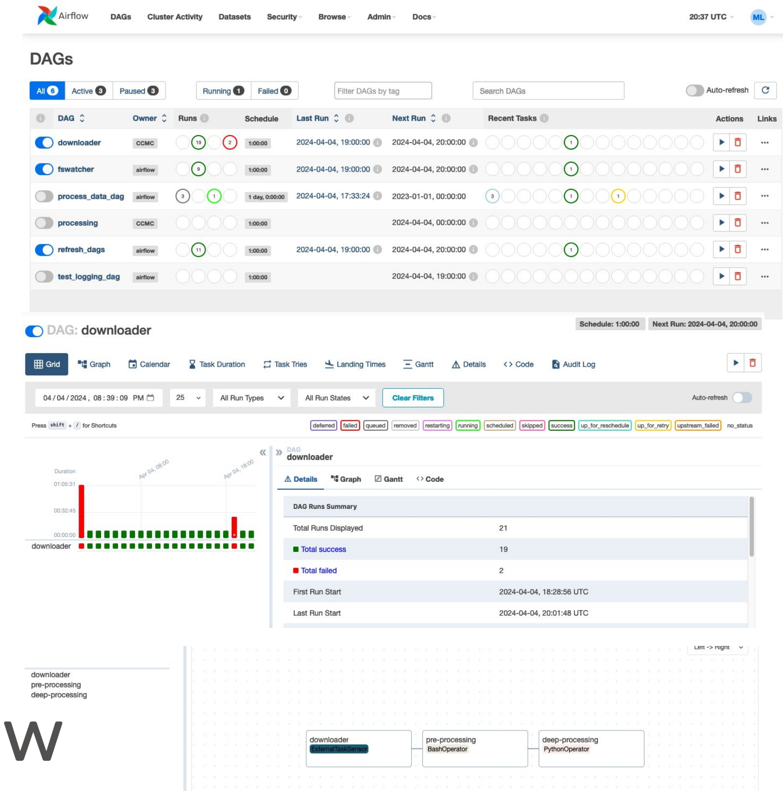
Cold Data
Sporadically Accessed

Role of Apache Airflow

Key Benefits

- Distributed workflow orchestration tool used by many organizations like AirBnb, Uber, Department of Energy Labs.
- Agents can run within servers and containers.
- Centralizes and schedules jobs across environments, which can typically be Python but supports other types depending on the Operator used.
- Efficient detection and alerting for pipeline failures.
- Allows non-airflow experts to manage, start, and restart the pipelines they own.

Different Roles for Different Folks



DAGs

All 6 | Active 3 | Paused 3 | Running 1 | Failed 0 | Filter DAGs by tag | Search DAGs | Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
downloader	CCMC	19 2	1:00:00	2024-04-04, 19:00:00	2024-04-04, 20:00:00	1	▶ 🗑	···
fswatcher	airflow	9	1:00:00	2024-04-04, 19:00:00	2024-04-04, 20:00:00	1	▶ 🗑	···
process_data_dag	airflow	3 1	1 day, 0:00:00	2024-04-04, 17:33:24	2023-01-01, 00:00:00	3 1 1	▶ 🗑	···
processing	CCMC		1:00:00		2024-04-04, 00:00:00		▶ 🗑	···
refresh_dags	airflow	11	1:00:00	2024-04-04, 19:00:00	2024-04-04, 20:00:00	1	▶ 🗑	···
test_logging_dag	airflow		1:00:00		2024-04-04, 19:00:00		▶ 🗑	···

DAG: downloader

Schedule: 1:00:00 | Next Run: 2024-04-04, 20:00:00

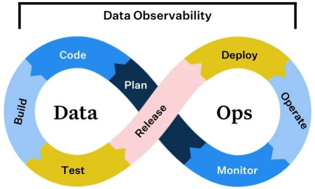
04 / 04 / 2024 , 08 : 39 : 09 PM | 25 | All Run Types | All Run States | Clear Filters | Auto-refresh

Press shift + / for Shortcuts

deferred failed queued removed restarting running scheduled skipped success up_for_reschedule up_for_retry upstream_failed no_status

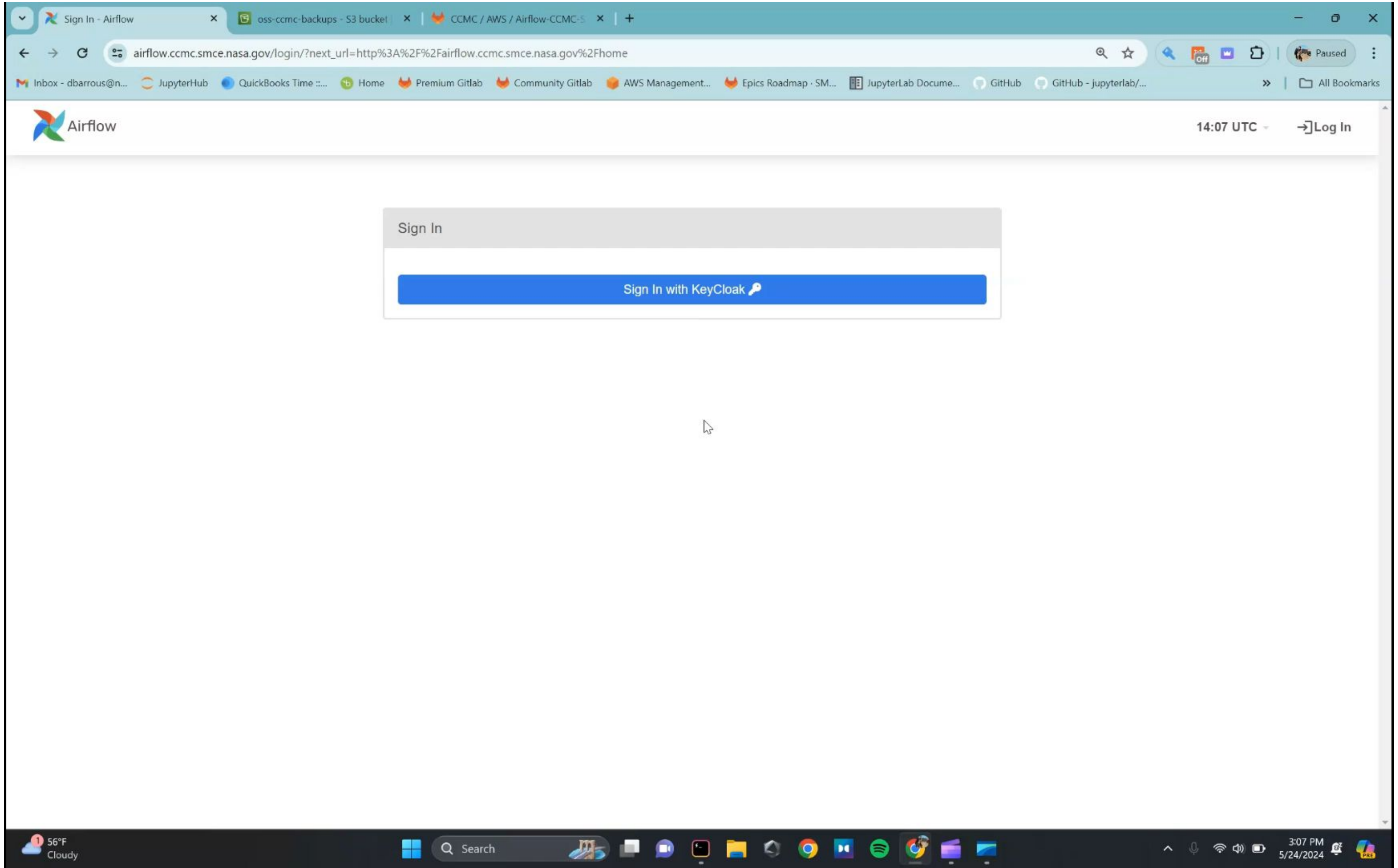
DAG
downloader

DAG Runs Summary	
Total Runs Displayed	21
🟩 Total success	19
🟥 Total failed	2
First Run Start	2024-04-04, 18:28:56 UTC
Last Run Start	2024-04-04, 20:01:48 UTC



Apache Airflow

CCMC Implementation of Airflow Demo

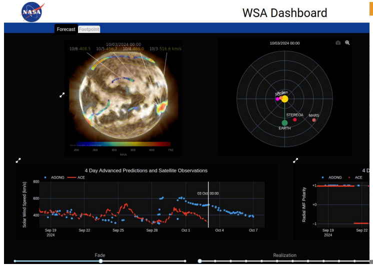


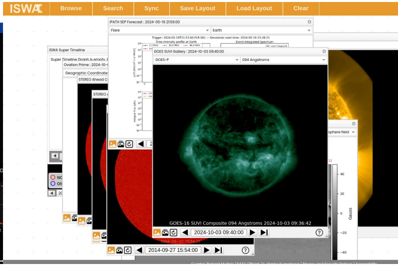
Sign In

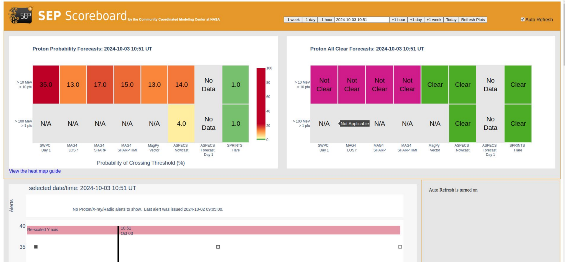
Sign In with KeyCloak

Improves our interconnected networks of simulation services and run result archives

- Enhanced **monitoring and error detection** for improved reliability.
- **Standardized metadata** capture to build a metadata database for tiering and improved DataOps.
- Support for complex workflows and data transformations.
- **Improved data reliability and accessibility** for researchers and forecasters for all of the services we offer.







Proton Probability Forecasts: 2024-10-03 10:51 UT

	SWPC Day 1	MAG4 LOS r	MAG4 SHARP	MAG4 SHARP HMI	MagPy Vector	ASPECS Nowcast	ASPECS Forecast Day 1	SPRINTS Flare
> 10 MeV > 10 pfu	35.0	13.0	17.0	15.0	13.0	14.0	No Data	1.0
> 100 MeV > 1 pfu	N/A	N/A	N/A	N/A	N/A	4.0	No Data	1.0

Probability of Crossing Threshold (%)

View the heat map guide

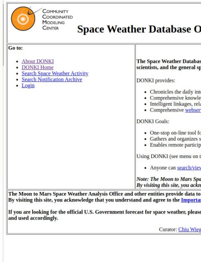
Proton All Clear Forecasts: 2024-10-03 10:51 UT

	SWPC Day 1	MAG4 LOS r	MAG4 SHARP	MAG4 SHARP HMI	MagPy Vector	ASPECS Nowcast	ASPECS Forecast Day 1	SPRINTS Flare
> 10 MeV > 10 pfu	Not Clear	Not Clear	Not Clear	Not Clear	Clear	Clear	No Data	Clear
> 100 MeV > 1 pfu	N/A	Not Applicable	N/A	N/A	N/A	Clear	No Data	Clear

selected date/time: 2024-10-03 10:51 UT

No Proton/X-ray/Radio alerts to show. Last alert was issued 2024-10-02 09:05:00.

Auto Refresh is turned on



Space Weather Database O

Go to:

- About DONKI
- DONKI Home
- Search Space Weather Activity
- Search Notification Archive
- Login

The Space Weather Databas
scientists, and the general sp

DONKI provides:

- Chronicles the daily int
- Comprehensive knowle
- Intelligent linkages, rela
- Comprehensive webser

DONKI Goals:

- One-stop on-line tool fo
- Gathers and organizes s
- Enables remote particip

Using DONKI (see menu on t

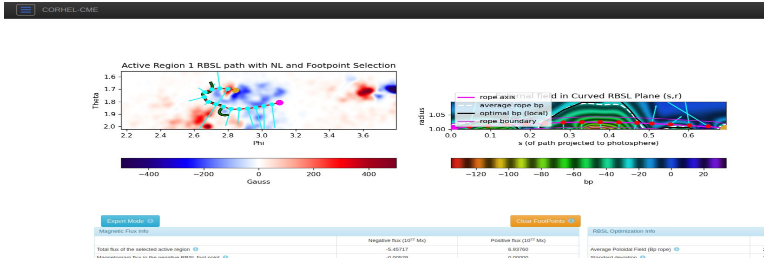
- Anyone can search/view

Note: The Moon to Mars Spa
By visiting this site, you ackn

The Moon to Mars Space Weather Analysis Office and other entities provide data to
By visiting this site, you acknowledge that you understand and agree to the Importa

If you are looking for the official U.S. Government forecast for space weather, please
and used accordingly.

Curator: Chiu Wieg



Magnetic Flux Info

	Negative flux (10^{22} Mx)	Positive flux (10^{22} Mx)
Total flux of the selected active region	-5.45717	6.93760
Magnetogram flux in the negative RBSL foot point	-0.00529	0.00000

RBSL Optimization Info

Average Poloidal Field (Bp rope)

Standard deviation

Leading Open Science at NASA: CCMC's Commitment

- Pioneering the use of AWS since 2019 to build collaborative, distributed systems.
- Open-sourcing tools like **Kamodo** via NASA's open science initiative.
- Actively contributing to global open-source projects such as **OpenSpace**.

With this in mind, we're looking at how to share our distributed Airflow setup to help others facing similar challenges.

Reach out to us if interested!

Future Outlook and Conclusion

- Continue migrating more jobs and file transfers to Airflow; collaborate with researchers to help them create and manage their own workflows.
- Develop a metadata database to enhance data captured through Airflow.
- Explore additional ways to improve the setup; Airflow remains a flexible tool within a centralized distributed workflow management system.
- Overall, Airflow has met our needs at CCMC and will be a key component in developing interconnected networks of simulation services and run result archives.

Thank you for your time!

Any Questions?

COMMUNITY COORDINATED MODELING CENTER

For more information on CCMC, please visit our website:

Contact Email: **dbarrous@navteca.com**

Goddard Space Flight Center