Application for SFSU Research Technician 1 (Computer Science) Position

Research Skills, Experience, Background, and Interests
Mark Yashar
http://www.linkedin.com/in/markyashar
https://github.com/markyashar
January 12, 2017

Outline

- Broad overview of research interests, background, and relevant scientific/technical skills, including academic coursework and additional training
- Specific examples of research experience and background, with some focus on analytical, numeric, and/or algorithmic work
  - Meteorological and CO_2 Regional Modeling (UCB)
  - Square Kilometer Array (SKA) research & development (UIUC)
  - Dark Energy Research (UCD)
- Summary

Overview of Research Background and Scientific/Technical skills

Interested in utilization and development of modeling, data analysis, statistical, mining, reduction, and processing algorithms, software, methods, and techniques in a wide range of possible scientific, scientific computing, and engineering disciplines. Examples:

- Monte Carlo methods and techniques (e.g., for parameter and error estimation)
  - Markov Chain Monte Carlo (MCMC) (including Bayesian analysis) and Metropolis Hastings algorithms
- Data and image handling, management, reduction, processing and analysis
- Data visualization and associated software tools

Overview of Research Interests and Scientific/Technical Skills

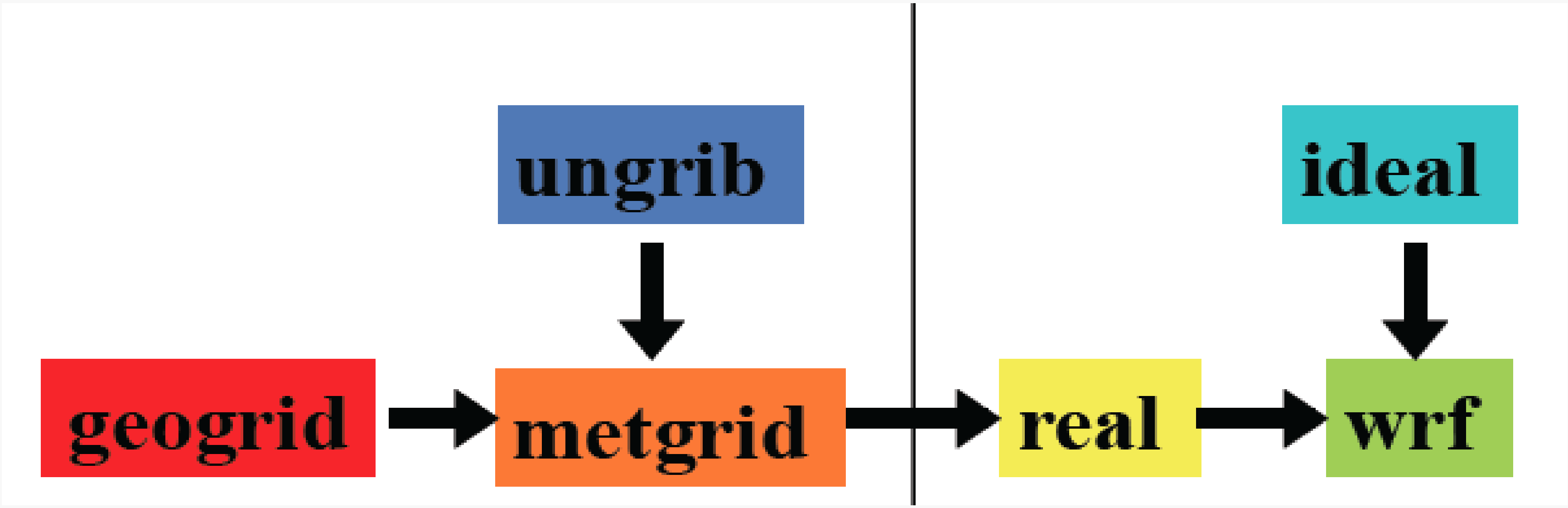
- Statistics and error analysis (e.g., propagation of errors, Gaussian distributions, covariance, and **correlation methods**)
- Modeling, simulation and model fitting (i.e., optimization)
  – Chi-square fitting
  – Nonlinear models
  – Confidence limits on estimated model parameters
  – Machine learning algorithms (e.g., Linear regression and logistic regression)
- Scientific/technical writing

Examples of Research and Technical Experience and Interests

Meteorological & CO_2 Regional Modeling (Supv.: I. Fung, UCB)

- Carried out research focused on mesoscale and regional atmospheric transport modeling and analysis of anthropogenic and biogenic CO_2 emissions for northern California for mutli-scale estimation and quantification of atmospheric CO_2 concentrations.
- Utilized **Weather Research and Forecasting (WRF)** model software which included use of **Runge-Kutta** methods for solving ODE's and use of **Euler** and **Navier-Stokes equations** routines for computing pressure gradient, buoyancy, horizontal and vertical diffusion, wind speed, etc.

WPS and WRF Program Flow (Dudhia, 2013)



Examples of Research Experience and Interests

Meteorological & CO_2 Regional Modeling (Supv.: I. Fung,UCB)

- Made extensive use of **WRF** (written mostly in **FORTRAN 90**), **WRF-Chem** coupled weather-air quality model for atmospheric (passive tracer) transport simulations, and **WRF-VPRM** biospheric model to simulate CO_2 biosphere fluxes and atmospheric CO_2 concentrations.
  - Involved use of **R statistical scripting language**, **NCL**, MATLAB, **Python (numpy, netCDF4), Ferret,** and **NCO tools** for additional pre- and post-processing, modification, visualization and analysis of **netCDF** files and output **diagnostic variables** of simulation results (e.g., pressure, temperature, relative humidity)

Examples of Research Experience and Interests

Meteorological & CO_2 Regional Modeling (Supv.: I. Fung, UCB)

- Troubleshooted and debugged **WRF**, **WRF-Chem**, and **VPRM** simulation runs and results to increase data efficiency and to decrease time to solve problems
- Installed, compiled, built, configured, and ran WRF, WRF-Chem, and VPRM on **NERSC multi-core supercomputing system** ("Hopper"; Cray XT CLE/Linux x86_64, PGI compiler w/gcc, distributed memory parallel mode, **MPI**) and submitted batch job scripts to this system to run code in parallel on compute nodes via **MPI** tasks
- Assisted students and post-docs in installing, configuring, and running WRF and WRF-Chem software on "Hopper"

Examples of Research Experience and Interests

Meteorological & CO_2 Regional Modeling: WRF preprocessing, post-processing, analysis, and visualization

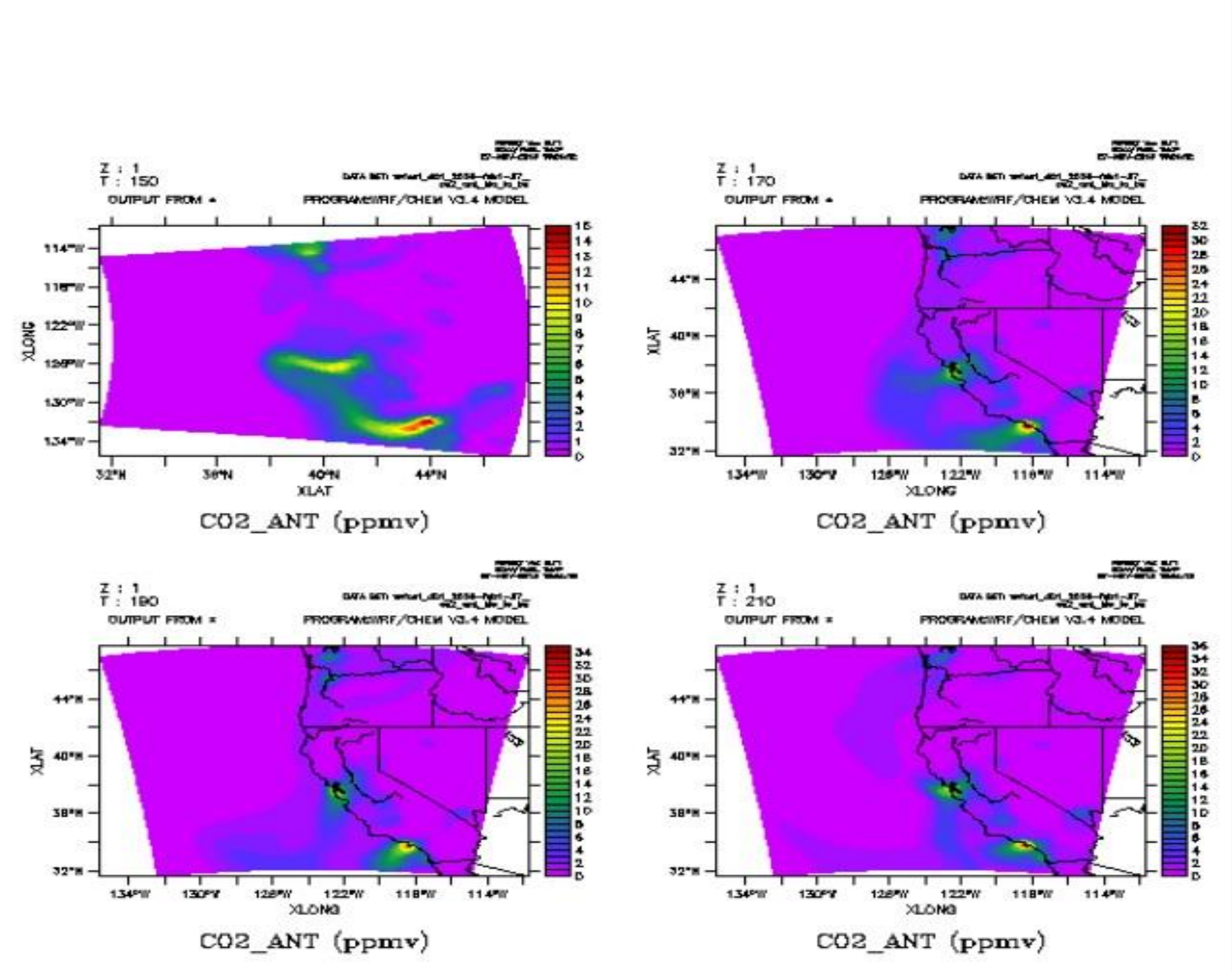
- WRF Preprocessing
  - Modified **FORTRAN 90** code to convert (input) CO2 flux emissions inventory data files between **HDF5**, binary, and **NetCDF-4** formats to enable and expedite further processing
  - Modified and utilized R scripts to edit **NetCDF-4** files to be in correct format and containing necessary CO2 tracer emission fields and data to be input to WRF simulation runs

Examples of Research and Technical Experience and Interests

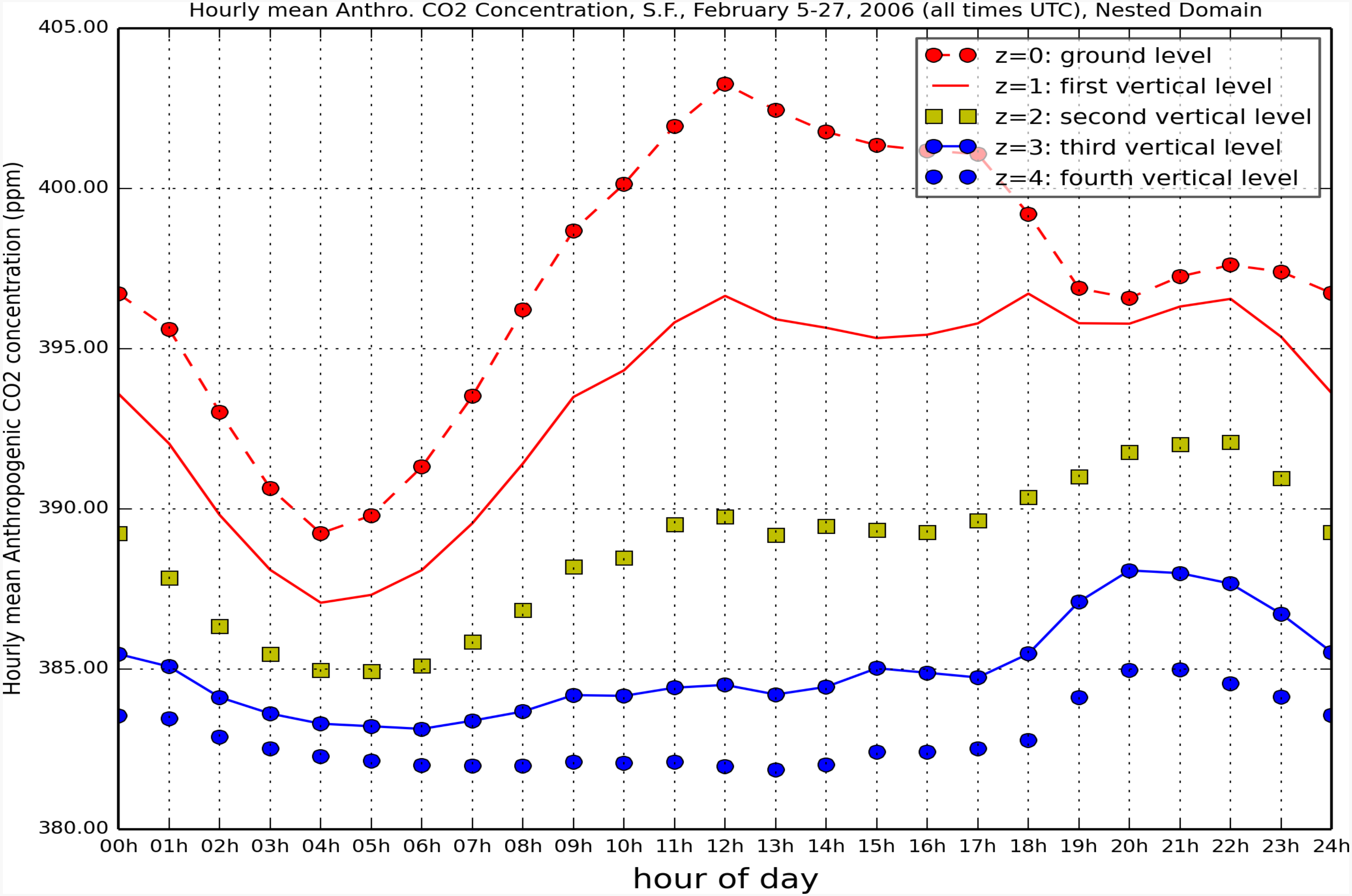
Meteorological & CO_2 Regional Modeling: WRF preprocessing, post-processing, analysis, and visualization

- WRF visualization and analysis tasks (post-processing)
  - Wrote and utilized **NCL, Ferret, and Python (numpy)** scripts and NCO package (e.g., ‘ncdump’, ‘ncview’) to visualize and analyze WRF model output **diagnostic variables** (e.g., CO2 concentrations, pressure, temperature, wind vectors) in form of contour maps, time series plots, and animations
  - Used **NCL, NCO, Ferret, and Python** to compute summary statistics of model variables (e.g., means)

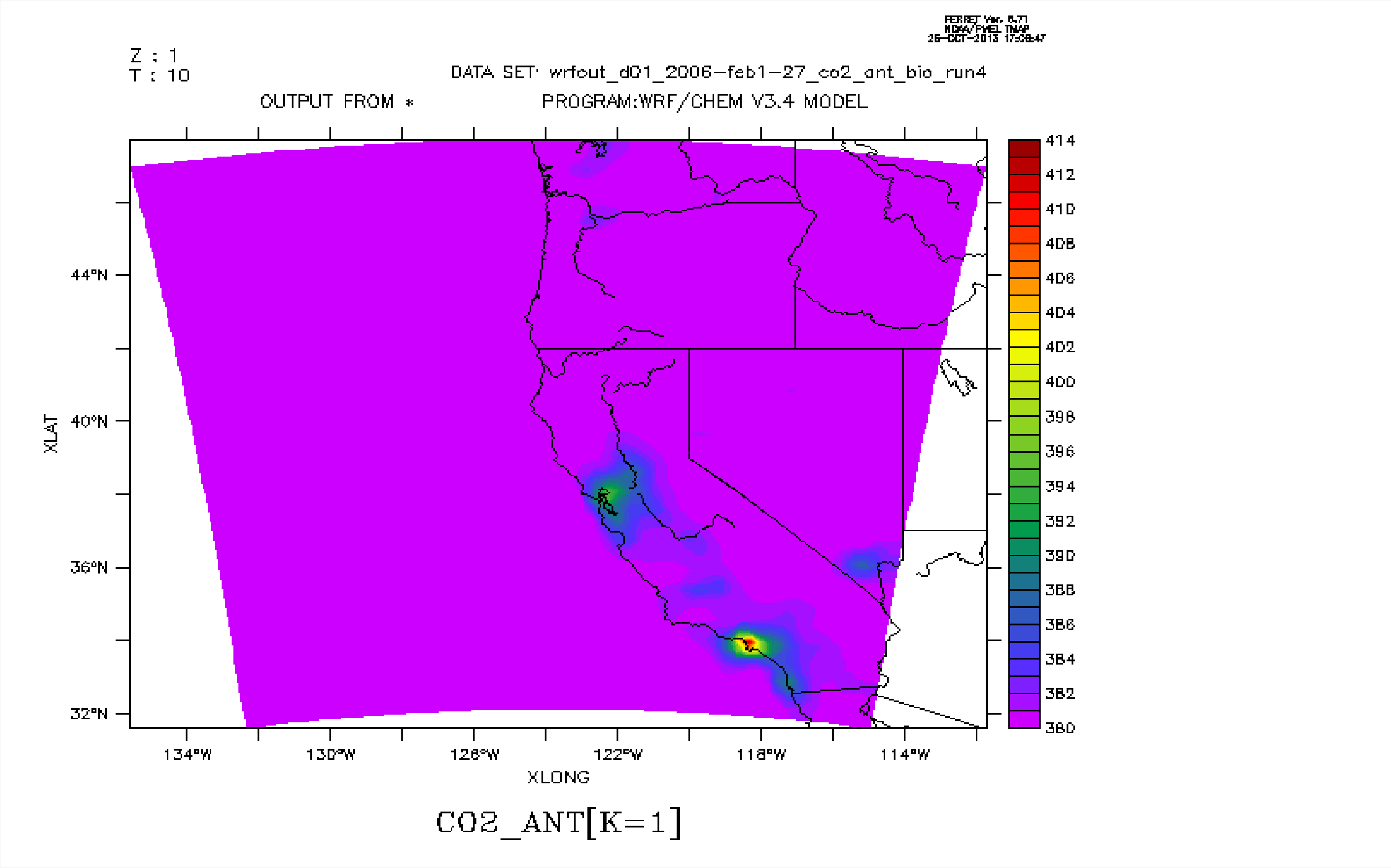
WRF Simulation Results: Contour Maps



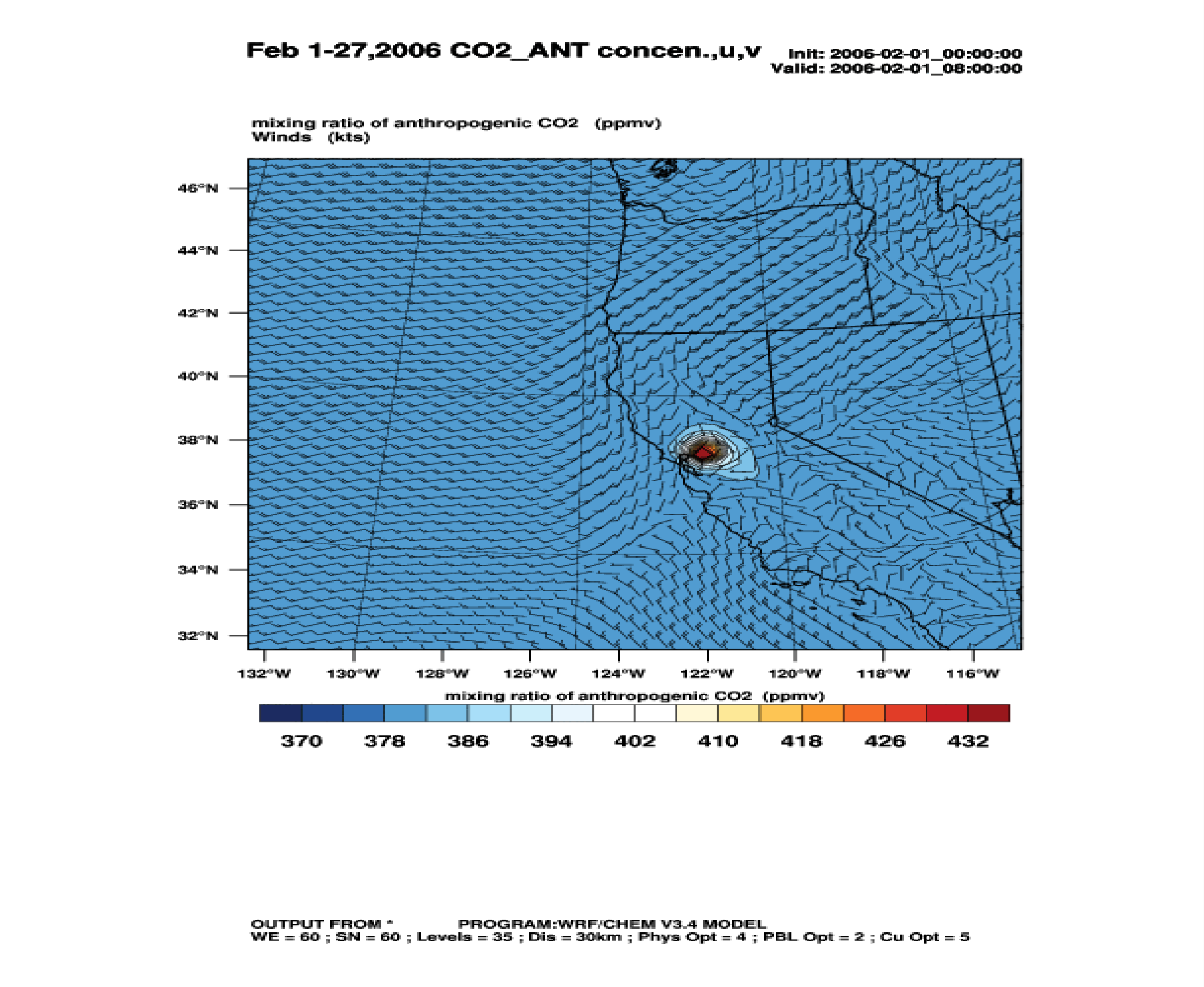
WRF Simulation Results: Time Series Plots



WRF Simulation Results: Animations (Ferret)



WRF Simulation Results: Animations (NCL)



Examples of Research Experience and Interests

SKA Research and Development (Supv.: A. Kemball, UIUC)

- R&D in SKA calibration and processing algorithms and computing with a focus on cost and feasibility studies of radio imaging algorithms and direction-dependent calibration errors
  - Evaluated the computational costs of non-deconvolved images of a number of existing radio interferometry algorithms used to deal with non-coplanar baselines in wide field radio interferometry and co-authored a corresponding internal technical report with A. Kemball

    (**github.com/markyashar/Job_Application_Materials/cpg_memo3_v1.1.pdf**)

Examples of Research Experience and Interests

SKA Research and Development (Supv.: A. Kemball, UIUC)

Starting point for these computations: The computational costs depend on data rate, computational intensity, and efficiency of parallel computing implementation, η, by way of following equation:

$$\textit{computational cost} = (\textit{data rate}) \times (\textit{computational intensity}) \times \eta^{-1}$$

where units of computational intensity are in operations per byte, and

$$\textit{computational intensity} = (\textit{total number of operations}) \,/\, (\textit{number of input and output points})$$

where number of input and output data points is essentially same as number of data transmitted.

Basic Imaging Considerations: Fourier Synthesis Imaging

- Generating a sky image with an interferometer array translates into reconstructing sky brightness I at a given frequency and as function of angular coordinates (l,m) on sky using measured visibility data V(u,v,w)
- For narrow field imaging, in which all the baselines are in same (u,v) plane and w=0, V(u,v) and I(l,m) are a 2D Fourier pair, with the normalized primary beam pattern of the antenna A(l,m) factored in
- Therefore, can recover brightness I(l,m) from a 2-D inverse FT of V(u,v) and deconvolution process

Fourier Synthesis Imaging: Wide-field, 3-D Imaging

- For more general case of non-coplanar baselines, in which $w \neq 0$, 2-D FT cannot be used. This is case for SKA for wide-field, low-frequency imaging will constitute large part of science case
- In 3-D imaging, 3-D FT of sampled visibility defines a 3-D "dirty" image $\mathbf{I_3^D}$ **(l,m,n)** and a 3-D PSF $\mathbf{B^{3D}}$ **(l,m,n),** which is convolved with 3- D image cube **F(l,m,n)**, where $\mathbf{I_3^D = F*B^{3D}}$. Basis of 3-D imaging method is to form I_3^D (l,m,n) and $\mathbf{B^{3D}}$ **(l,m,n)** and then to solve for **F(l,m,n)** using any number of deconvolution algorithms
- Focused on computational loads and costs of number of imaging algorithms that dealt with 3-D imaging, and non-coplanar baselines effect (wide-field imaging)

Examples of Research and Technical Experience and Interests

SKA Research and Development (Supv.: A. Kemball, UIUC)

- **M. Yashar**, A. Kemball, Computational Costs of Radio Imaging Algorithms Dealing with the Non-coplanar Baselines Effect:I, TDP Calibration and Processing Group Memo #3

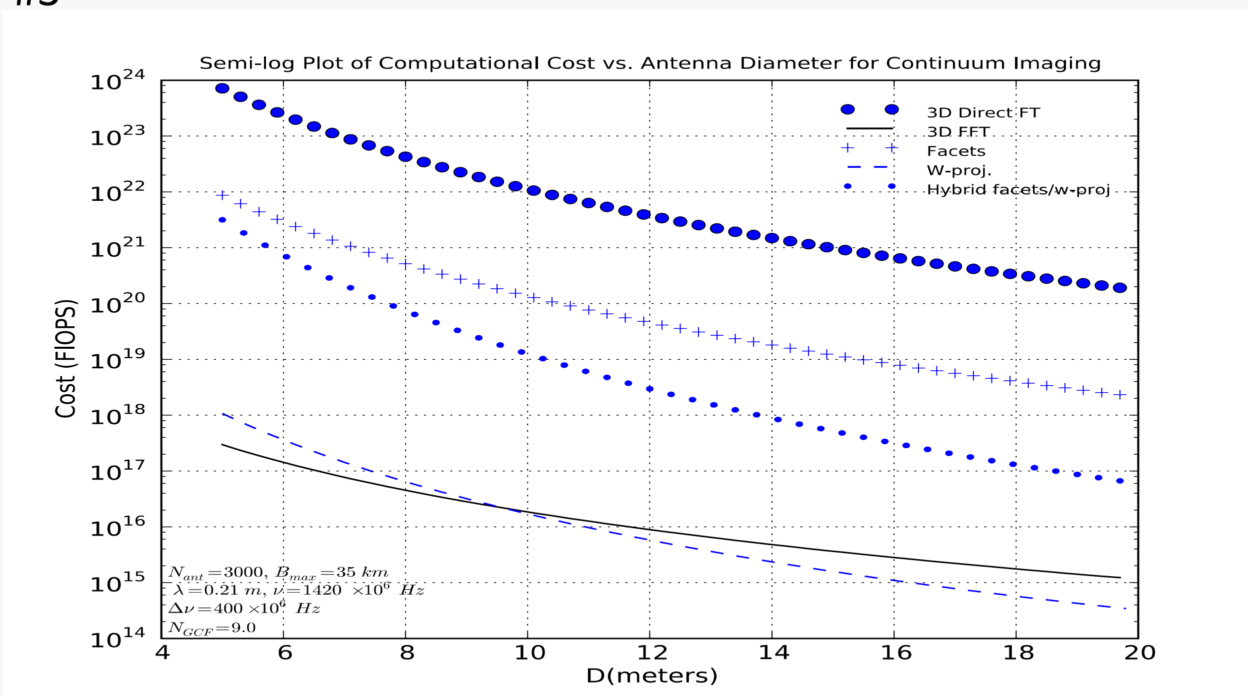


Figure 1: Semi-log y plots of computational costs (without consideration of deconvolution and parallel computing efficiency η) vs. antenna diameter D for continuum imaging for the 3-D direct FT, 3-D FFT, facets, w-projection, and hybrid facets/w-projection imaging algorithms.

Examples of Research and Technical Experience and Interests

SKA Research and Development (Supv.: A. Kemball, UIUC)

- **Simplified Example:** Computational Costs of DFT vs. FFT
  - Discrete Fourier Transform (DFT) is discrete version of FT which transforms a signal (discrete series) from time domain representation to frequency domain representation
  - Fast Fourier Transform (FFT) is an efficient algorithm for calculation of DFT
  - Computing a DFT of N points takes, e.g., **$O(N^2)$** time or operations, while FFT can compute same result in only **$O(N*Log(N))$** time or operations, which is a substantial gain (and therefore decrease in computational costs) for large sequences.
  - For imaging applications, N could be number of image pixels along the l, m, n axes

Examples of Research and Technical Experience and Interests

SKA Research and Development (Supv.: A. Kemball, UIUC)

- Implemented numerical and imaging simulations (Meqtrees, CASA – **Python,C++**) and Monte Carlo simulations (**Python--numpy**) to address cost, feasibility, dynamic range, and image fidelity issues related to calibration and processing for SKA and dependence of these issues on key antenna and feed design parameters (e.g., sidelobe level, mount type). Co-authored corresponding technical memo: A. Kemball, T. Cornwell, **M. Yashar**, Calibration and Processing Constraints on Antenna and Feed Designs for SKA: I, TDP Calibration and Processing Group Memo #4
  - Installed, built, compiled, configured **C++ software development (OOP)** environment for CASA (gdb, ddd, Eclipse debuggers, GNU tools) and made modifications to C++ code to carry out simulations and test imaging algorithms

Examples of Research and Technical Experience and Interests

Dark Energy Research (Supv.: A. Albrecht, UCD)

- MCMC analysis (involving extensive use of MATLAB code) of dark energy quintessence model (IPL or Ratra-Peebles) that included utilization of DETF data models that simulated current and future data sets from new and proposed observational programs (e.g., SKA observations)
- Wrote and submitted batch job scripts to run MATLAB MCMC code on Linux computing cluster to expedite running of MCMC simulations and generation of MCMC output
- Troubleshooted and debugged MATLAB code, simulation runs and results

Markov Chain Monte Carlo

- Technique often applied to solve **optimization problems** (e.g., parameter fitting) in large-dimensional spaces
  - MCMC simulates likelihood surface for set of parameters by sampling from posterior distribution via series of random draws
  - Chain steps semi-stochastically in parameter space via, e.g., **Metropolis-Hastings algorithm**, such that more probable values of parameter space are stepped to more often
  - Number of different convergence tests and criteria are used to ascertain whether MC does good job of reflecting posterior distribution of the model

Markov Chain Monte Carlo

- In this way, can analyze parameter space of IPL quintessence in light of DETF data models and evaluate likelihood function of parameters of our model
- Once Markov chains of our models in parameter space have been computed, can extract likelihood contours from distribution of models and display them as projected 2D likelihood contour plots.
- This then gives us picture of shape of likelihood region of all the parameters in our models in the whole multidimensional parameter space if we were to plot likelihood contours for each pair of parameters in the parameter space

Dark Energy Research (Additional Details)

- Generated simulated data sets for Lambda-CDM background cosmology as well as case where dark energy was provided by specific IPL model.
- Used MCMC algorithm (also including Bayesian, and various mathematical modeling and uncertainty quantification methods/skills) to map likelihood around each fiducial model and moving to succession of random point in space using Metropolis-Hastings stepping algorithm.
- From associated likelihood contours, found that respective increase in constraining power with higher quality data sets produced by analysis gave results that were broadly consistent with DETF parameterization that they used.

Examples of Research and Technical Experience and Interests

Dark Energy Research (Supv.: A. Albrecht, UCD)

- Lead author of paper published in Physical Review D on research results: **M. Yashar**, B. Bozek, A. Albrecht, A. Abrahamse, M. Barnard, Exploring Parameter Constraints on Quintessential Dark Energy: The Inverse Power Law Model, Physical Review D, 79, 103004, 2009.
  - From the associated likelihood contours, found that the respective increase in constraining power with higher quality data sets produced by analysis gave results that were broadly consistent with the DETF for the dark energy parameterization that they used. Also found, consistent with other findings, that for a universe containing dark energy described by the IPL potential, a cosmological constant can be excluded by high quality “Stage 4” experiments by well over 3σ.

Examples of Research and Technical Experience and Interests

Dark Energy Research (Supv.: A. Albrecht, UCD)

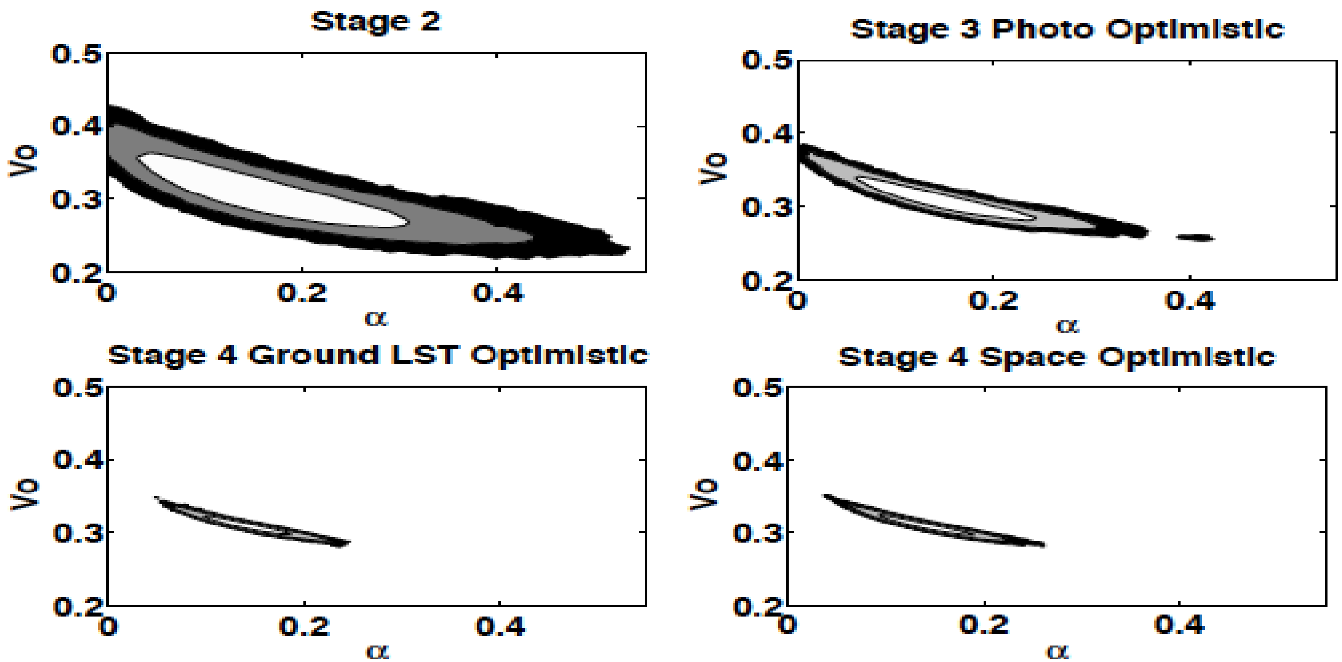


FIG. 9: $V_0-\alpha$ 1σ (68.27%), 2σ (95.44%) and 3σ (99.73%) likelihood contours for DETF optimistic combined data sets generated from a selected IPL background cosmological model.

Examples of Research and Technical Experience and Interests

Dark Energy Research (Supv.: A. Albrecht, UCD)

- Assisted graduate student in generating 3-D χ^2 plots with MATLAB that helped develop intuition into actual physical behavior of Albrecht-Skordis (AS) dark energy quintessence model – better understanding than would have been allowed by running full MCMC on larger parameter space
  - This systematic investigation revealed some numerical problems and issues in student's analysis of AS model.

Summary

- Overview of research interests and scientific/technical skills
- Examples of research experience and interests
  - Meteorological and CO_2 regional modeling
  - Square Kilometer Array (SKA) research & development
  - Dark Energy research

Appendix -- References

- SKA R&D Technical Memo: https://github.com/markyashar/Job_Application_Materials/cpg_memo3_v1.1.pdf
- Additional write-up of SKA R&D work: https://github.com/markyashar/SKA_NCSA_Research_Simulations/Final_Report_SKA_meqtrees_casa.pdf
- Dark Energy Paper: https://github.com/markyashar/Job_Application_Materials/IPL_DE_Paper.pdf
- WRF and WRF-Chem Notes and Documentation:

  https://github.com/markyashar/UCB_WRF_Chem_Documents

Thank You

- Look forward to
  – helping you solve problems, make a contribution, and add value to the research group
  – acquiring new computational and analysis skills as necessary (and come up to speed on things quickly)

Thank you. Questions?

Extra Slides: Additional Technical Details

Example Deconvolution Algorithm: CLEAN

- "CLEAN" algorithm is computational algorithm to perform deconvolution on images created in radio astronomy
  - Algorithm assumes that image consists of a number of point sources
  - Iteratively finds highest value in image and subtract small gain of this point source convolved with PSF ("dirty beam") of observation, until highest value is smaller than threshold

Markov Chain Monte Carlo (Additional Details)

- For each step in MCMC chain, we integrate numerically to calculate theoretical quantities dependent on Dark Energy
  - Start integration at early times in the universe and end calculation in more recent times
  - Compare these values with the observables generated based on our fiducial models (via χ^2 calculated for each step of the MCMC chain)
  - With the uncertainties in the data forecast by DETF we can calculate likelihood for each step in the chain
  - When chain has converged, it's considered "fair sample" of posterior distribution, and density of points represents likelihood surface.

Metropolis Hastings Algorithm(Additional Details)

- **Metropolis-Hastings** is a random walk MC algorithm that generates a random walk using a proposal density and method for rejecting some of the proposed moves
  - Chain starts at arbitrary position in parameter space
  - Candidate position **ϴ’** for next step in chain is drawn from a proposal density **q(ϴ,ϴ’)**
  - Candidate point in parameter space is accepted and becomes the next step in chain with probability **α(ϴ,ϴ’)**
  - If proposal step **ϴ’** is rejected, point **ϴ** becomes next step in chain. Although many distributions are viable for proposed density **q(ϴ,ϴ’)**, for simplicity we’ve chosen (in DE example) to use Gaussian normal distribution

Metropolis Hastings Algorithm(Additional Details)

- Power of MCMC procedure lies in fact that it can probe posterior distributions that are quite different from proposal density **q(Ɵ,Ɵ)**
- For results of Markov chain to be valid, it must equilibriate, i.e., converge to stationary distribution
- If such a distribution exists, Metropolis-Hastings algorithm guarantees that chain will converge as chain length $\rightarrow \infty$
- In practice, however, must work with chains of finite length. From standpoint of computational efficiency, the shorter the chains can be and still reflect the posterior distribution of the parameters, the better