# High Efficiency Implementation of *PC* and *PC stable* Algorithms Yields Three-Dimensional Graphs of Information Flow for the Earth' Atmosphere

Technical Report

Colorado State University

Department of Electrical and Computer Engineering

**Report Nr. CSU-ECE-2014-1**

**Imme Ebert-Uphoff*** **and Yi Deng†**

September 3, 2014

**Abstract**

Causal discovery algorithms have recently been applied to several climate applications. In particular, in prior work we have developed methods to recover pathways of interaction in the global climate system, using the classic *PC* algorithm. However, standard implementations of the *PC* algorithm cannot handle the large number of variables and temporal models required for this application. This technical report shows that a more efficient implementation of the *PC* algorithm can provide speed gains of a factor of 1,000 or more. This in turn enables us to calculate graphs of information flow with much higher resolution grids. Furthermore, we can now - for the first time ever - calculate information flow graphs that extend over three dimensions, i.e. rather than just including *one* layer of the planet's atmosphere we can now capture interactions across several height layers.

## 1 Introduction

Causal discovery seeks to discover potential cause-effect relationships from observational data. While used extensively for decades in disciplines such as social science and economics, causal discovery has only recently been used in climate science. Requirements of climate applications can be challenging for existing implementations: they often require using a large number of variables, distributed over large spatial regions, and require the use of temporal (rather than static) models, which further increases complexity.

In prior work [3, 4] we have applied causal discovery algorithms in climate science to find potential cause and effect relationships from observed atmospheric data. The key idea for this application is to interpret large-scale atmospheric dynamical processes as information flow around the globe and to use causal discovery to identify the pathways of information flow around the globe. Specifically, by introducing a discrete, equally-spaced grid around the earth, we can use the causal discovery algorithms to calculate "graphs of information flow", which show the flow of information (i.e. interactions) around the globe. Which dynamical process is tracked through this method depends on the type of atmospheric variable observed in the data (e.g. geopotential height) and the time scale used (e.g. daily vs. monthly data). For details of this application and the basic methodology, please see [4, 5].

*iebert@engr.colostate.edu, Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado, USA

†yi.deng@eas.gatech.edu, School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA

The causal discovery algorithm we used for this application is the classic *PC* algorithm [8, 9], extended to provide temporal models [1], and incorporating the improvements of the *PC stable* algorithm [2]. This approach yielded good results for low resolution planar graphs of information flow. However, once we wanted to move on to high resolution grids, we quickly hit the limits of standard implementations. For example, using a grid with $200$ points around the globe, and using $15$ time slices, the algorithm needs to handle a total of $15 \times 200 = 3,000$ variables, which on a MacBook Pro or a standard Linux computer already required about 4 days. We would like to drastically increase the grid to have 400, 800, or maybe even 2,000 nodes. Furthermore, atmospheric information flow occurs in 3D, not 2D, thus we also want to move on to a spatial grid (and thus spatial graph) including more than one height layer around the earth at the same time.

# 2 Faster Implementation

We tried publicly available implementations of the *PC* algorithm, namely *TETRAD* (implemented in Java), *BNT* (implemented in Matlab) and *pcalg* (implemented in R). None of them allowed us to move considerably beyond the $3,000$ variables already required for the temporal model (15 time slices) with a planar low level grid (200 points). We first considered other types of algorithms, such as score-based algorithms and Granger graphical models. However, we liked the overall properties of constraint-based algorithms - of which the *PC* algorithm is the best known example - namely we find them to be reliable and transparent, i.e. each step of the process easy to understand. Thus we decided to try coming up with a more efficient *implementation* of the *PC* algorithm, rather than switching to a different algorithm. At a later time we plan to still try other algorithms.

As mentioned above, the available implementations of *PC* we found were in Java, Matlab and R. While Matlab and R are prime environments for mathematical computing, they are not that good at actual number crunching. C is known to be much better suited for that purpose. We thus implement the algorithms (PC, PC stable and their temporal extension) in C, using the GNU scientific library. It turned out that once we used tens of thousands memory localization becomes a crucial issue, as illustrated in the following example. When we increased the number of variables to tens of thousands the algorithm suddenly took a tremendous amount of time (days) for the simple task of just calculating the correlation matrix from the sample data. (This task is done once at the very beginning of the algorithm, as preparation for PC.) We solved this problem by simply *transposing* the data matrix, which holds all the observed data for all variables. Originally the observed data of an individual variable was stored in a *column* of the data matrix (with *rows* representing *one sample* of all variables). By instead representing each variable as a *row* in the data matrix, we moved the values for each variable closer together in memory, thus reducing access time to those, and in turn reducing calculation time for the correlation matrix from days to hours. In summary, by using C and by careful implementation and optimization, we achieved a speed factor of $300$ over the Java/Matlab/R implementations. Thus calculation with $3,000$ variables was reduced from 4 days to 20 minutes.

In the next step we added multi-threading, speeding calculations by another factor of 4 or more for most multi-core computers, such as many standard PC or Mac computers. This already allowed us to calculate our first **spatial** graphs of information flow, using $400$ grid points per layer, up to $6$ height layers and $15$ times slices, i.e. requiring a total of $400 \times 6 \times 15 = 36,000$ variables in the *PC* algorithm. First results are shown in the following section.

# 3 Results - Spatial Graphs of Information Flow

This section shows results which - to the best of our knowledge - constitute the first spatial graphs of information flow around the globe ever obtained. We use the same methodology used in [4, 5], just that with our high-speed implementation we can now generate graphs with higher resolution and in three dimensions.

## 3.1 Data and Parameters Used

We use daily geopotential height data obtained from NCEP-NCAR reanalysis data [6, 7] for years 1950-2000. In some runs we used 4 layers of geopotential height (850mb, 500mb, 250mb, 50mb), in others we used six

height layers (925mb, 850mb, 500mb, 250mb, 50mb, 30mb)[1]. Furthermore, for each year only daily data for boreal winter is used (Dec-Jan-Feb). We use 400 geographical locations around the globe, and 15 temporal slices (of which we discard the first slices to overcome the initialization problem discussed in [4, 5]). Results are for PC stable, using $\alpha = 0.1$. For this experiment we chose to use for the edge directions only the temporal constraints. Therefore all edges for delay=1 or more days have directions (purely from the temporal constraints), while none of the edges for delay=0 have a direction, where the delay denotes the approximate time it takes for the signal to travel from cause to effect.

## 3.2 Figures

We provide two types of plots to show the atmospheric information flow in three dimensions: (1) Spherical plots provide a spatial image of the earth with connections from all height layers surrounding it (at their respective heights); and (2) Stereographic projection plots provide stereo-graphic projections of all height layers, stacked on top of each other. While the spherical plots are more intuitive to interpret and show information flow for both the Northern and Southern hemisphere in a single plot, it is hard to make out any details, in particular which connections belong to which height layers. Thus the stereographic projection plots are much more useful, even though they are less intuitive and show only connections for one hemisphere per plot.

To indicate connections between different layers, we use the following color code for the individual edges:

- Black: edge that is completely within a layer.
- Red: edge is going up in the atmosphere, i.e. from a lower height layer to a higher height layer.
- Green: edge is going down in the atmosphere.
- Green is also the default color, if we do not know whether an edge is going up or down, as is the case for all zero-delay edges that are not within a single height layer.

Figures 1 to 9 show results for a run with 4 height layers, while Figures 10 and 11 show results for a run with 6 height layers. For the 4 height layers, Figures 1 to 3 show the strongest connections that require about 0 days to travel from cause to effect (almost instantaneous). Figures 4 to 6 show the corresponding figures for connections spanning about 1 day, and Figures 7 to 9 for connections spanning about 2 days (very few connections span 2 days).

The most interesting plots for 4 layers are Figures 5 and 6, which show most clearly the directions of information flow for non-instantaneous connections. As one would expect the highest layer (at $50mb$) is relatively isolated from the other layers, with no strong interaction detected with any of the other layers (only black arrows). There is considerable interaction between the other three layers, and not all of it is exactly vertical. Much can be learned from plots of this type. In particular, we can generate this type of plot for different subsets of data, e.g. only using date ranges for certain atmospheric conditions. By comparing the resulting plots for those different conditions we can learn about specific information flow for each of those scenarios.

Figures 10 and 11 show results from using 6 height layers for strongest connections spanning 1 day. These figures illustrate that we can indeed run PC even with 36,000 variables, and remind us that more work needs to be done on studying the impact that the selection of height layers (i.e. which height layers are included) has on the results. For example, comparing the results in Fig. 10 to those in Fig. 5, we see that adding another height layer at the bottom (925mb) significantly reduces the number of arrows at the 850mb layer in Fig. 10. See also Section 4 which discusses future work.

## 3.3 Interpretation of Results

All the figures in this report were obtained from daily geopotential height data at different height layers and thus each arrow in these plots represents the pathway of *large-scale atmospheric waves* in three dimensions. Some of the information we can obtain from these plots are

[1] Note that a *higher* pressure means a *lower* layer, i.e. closer to the planet's surface, so for example 850mb is a lower layer than 500mb.

1. Location of the maximum wave source (largest number of upward pointing arrows).

2. Preferred pathways of wave propagation.

That type of information cannot be obtained from traditional methods, so represents new knowledge about the inner workings of our planet's climate. (Note also that no frequency filtering was used to obtain these results.) Being able to generate plots like the ones shown here is useful to better understand the effect of climate change, and to study selected dynamic processes.

# 4 Future work

We have only scratched the surface of the methods presented here, in terms of theory and interpretation, as well as their use to learn more about the internal workings of certain dynamic processes in the atmosphere. Some of the research to be addressed include:

- **Even higher efficiency implementation:** As illustrated in Section 2, optimization of local memory is a crucial speed factor for the implementation. So far we only localized the memory in a few places (e.g. transposing the data matrix), and believe that there additional adjustments throughout the code will result in significant additional speed-up. We are also planning to implement the message passing interface (MPI) so that we can use several nodes on a Cray simultaneously.

- **Selecting height layers:** As mentioned in Section 3.3, we need to further study the effects of selecting height layers on the results. We plan to develop guidelines for how to interpret the results based on the chosen heights, as well as potentially find ways to compensate for varying distances between the height layers in the causal discovery algorithms.

- **Applications in climate science:** Of course, we have only touched the surface of what we can do even with the current implementation. We can now generate information flow graphs for spatial grids and we will investigate what we can learn from them about our planet's climate using different types of atmospheric variables, different resolutions, etc.

- **Applications in bioinformatics:** This high-speed implementation of causal discovery may also be useful for applications in bioinformatics that seek to identify potential cause effect relationships between large numbers of variables. Sample applications include gene regulatory networks and finding neural connections in the brain.

# Acknowledgment

This work was supported in part by NSF Climate and Large-Scale Dynamics (CLD) program Grant AGS-1147601 awarded to Yi Deng.

# References

[1] T. Chu, D. Danks, and C. Glymour, *Data Driven Methods for Nonlinear Granger Causality: Climate Teleconnection Mechanisms*, Tech. Rep. CMU-PHIL-171, Dep. of Philos., Carnegie Mellon Univ., Pittsburgh, Pa., 2005.

[2] D. Colombo and M.H. Maathuis, *Order-independent constraint-based causal structure learning*, (arXiv:1211.3295v2), 2013.

[3] Y. Deng and I. Ebert-Uphoff, *Weakening of Atmospheric Information Flow in a Warming Climate in the Community Climate System Model*, Geophysical Research Letters, 7 pages, doi: 10.1002/2013GL058646, Jan 2014.

[4] I. Ebert-Uphoff and Y. Deng, *A New Type of Climate Network based on Probabilistic Graphical Models: Results of Boreal Winter versus Summer*, Geophysical Research Letters, vol. 39, L19701, 7 pages, doi:10.1029/2012GL053269, 2012.

[5] I. Ebert-Uphoff and Y. Deng, *Causal Discovery for Climate Research Using Graphical Models*, Journal of Climate, Vol. 25, No. 17, doi:10.1175/JCLI-D-11-00387.1, pp. 5648-5665, Sept 2012.

[6] E. Kalnay et al., *The NCEP / NCAR 40-year reanalysis project*, Bull. Am. Meteorol. Soc., 77, 437471, doi:10.1175/1520-0477 (1996)077¡0437:TNYRP¿2.0.CO;2, 1996.

[7] R. Kistler et al., *The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation*, Bull. Am. Meteorol. Soc., 82, 247267, doi:10.1175/1520-0477(2001)082¡0247:TNNYRM¿2.3.CO;2, 2001.

[8] P. Spirtes and C. Glymour, *An algorithm for fast recovery of sparse causal graphs*, Social Science Computer Review, 9(1):6772, 1991.

[9] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, Springer Lecture Notes in Statistics. 1st ed. Springer Verlag, 526 pp., 1993.

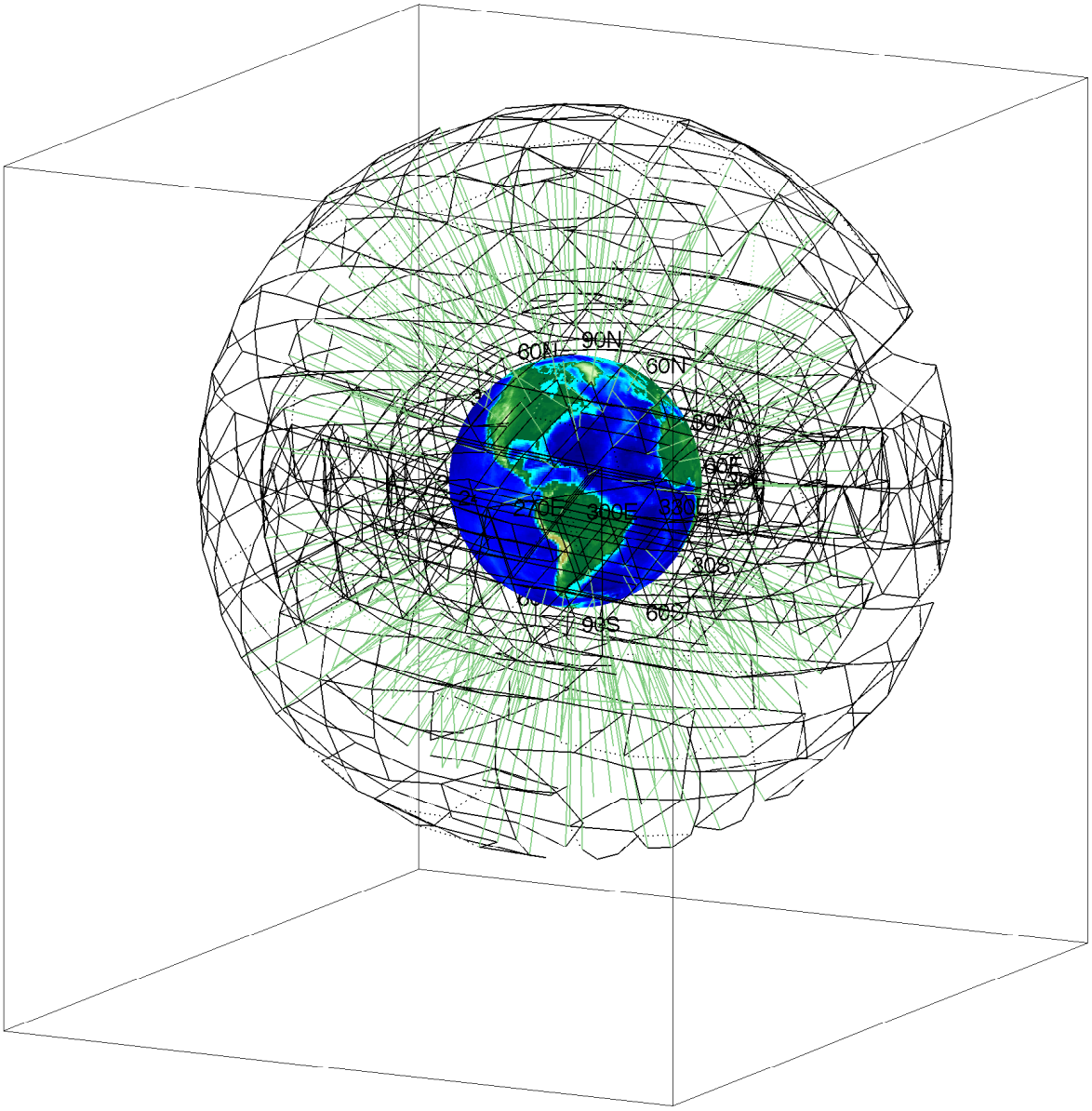


Figure 1: Spherical plot for 4 layers, strongest connections with travel time 0 days

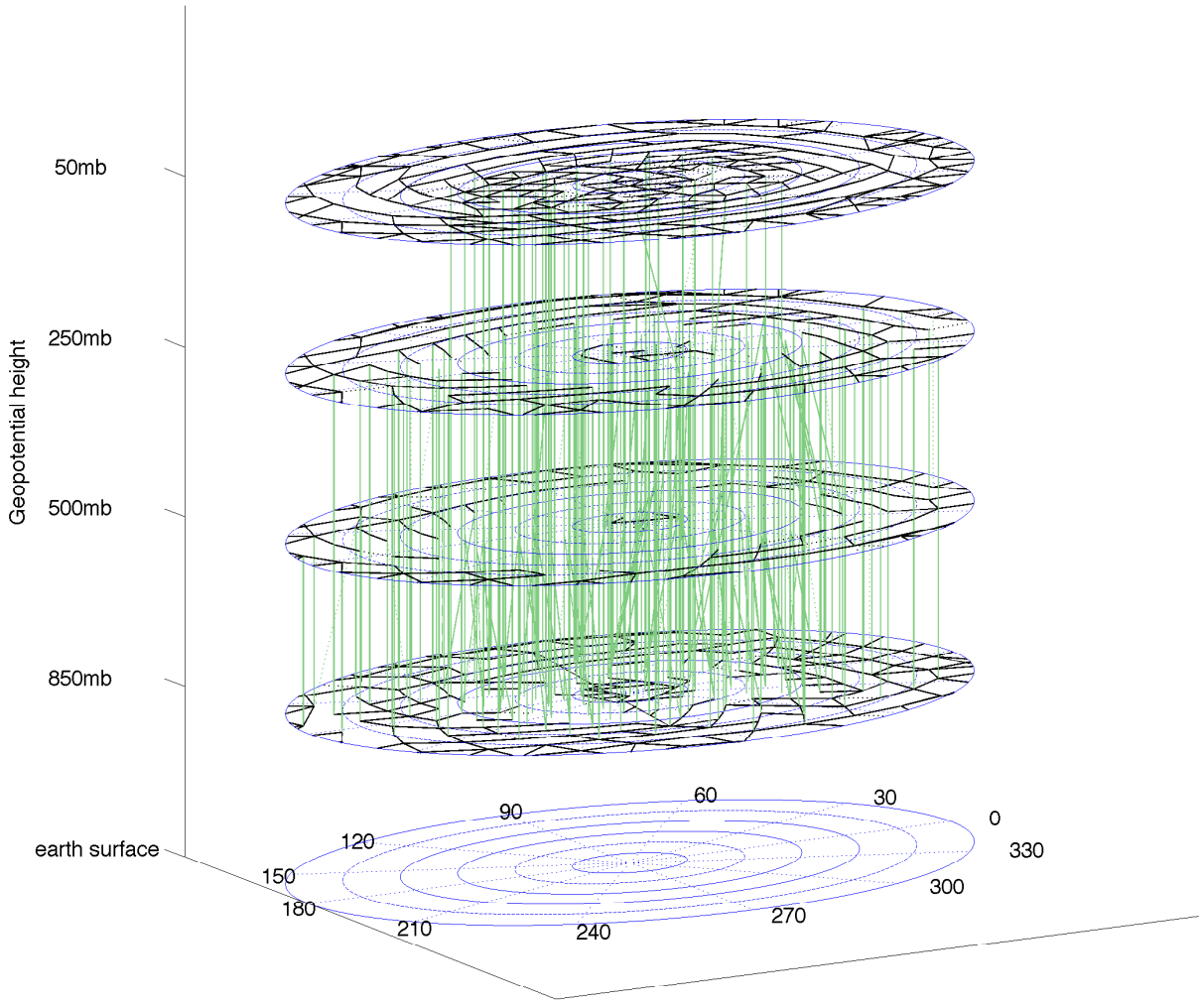


Figure 2: Stereographic projection plot for 4 layers, strongest connections with travel time 0 days, Northern hemisphere

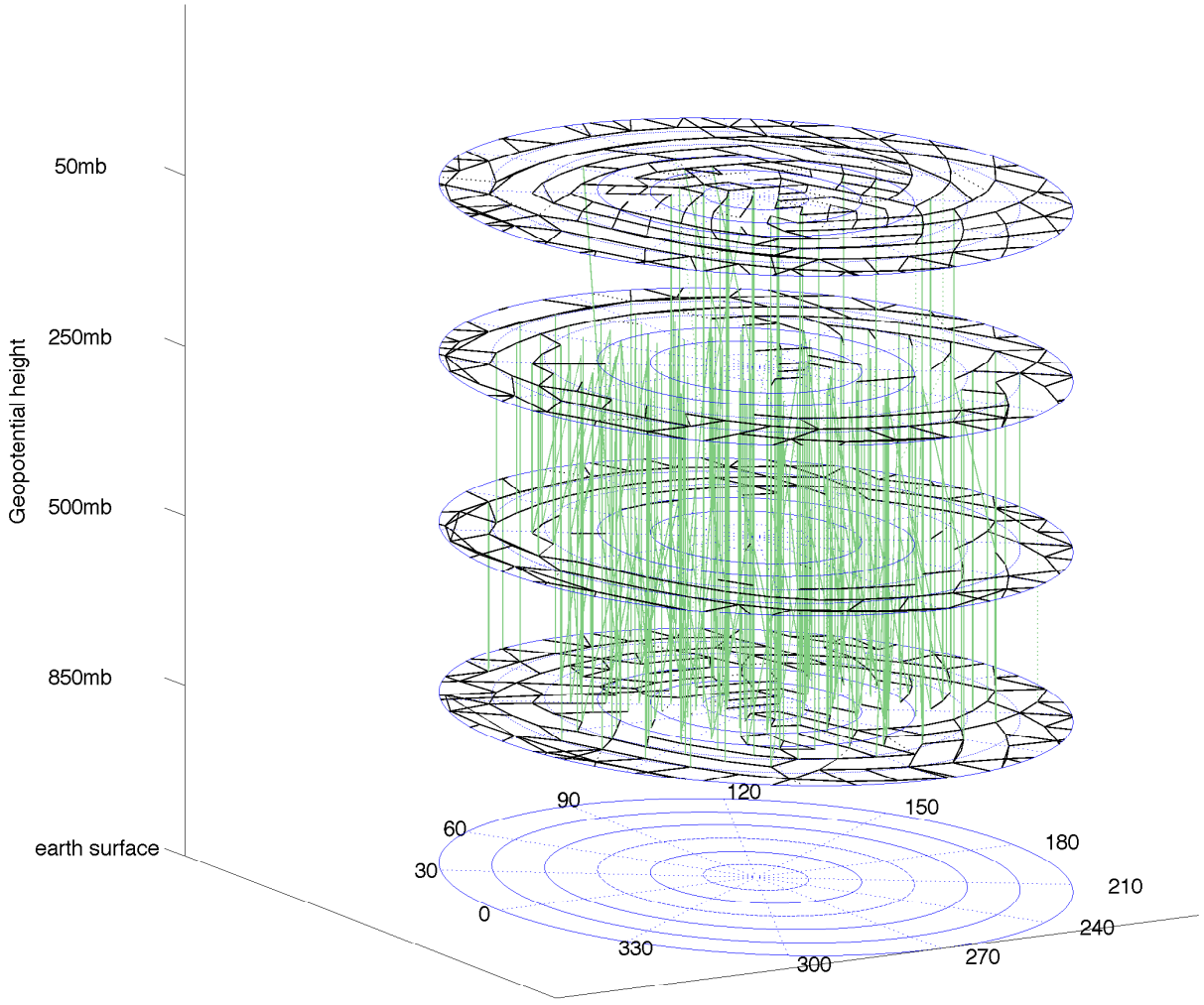


Figure 3: Stereographic projection plot for 4 layers, strongest connections with travel time 0 days, Southern hemisphere

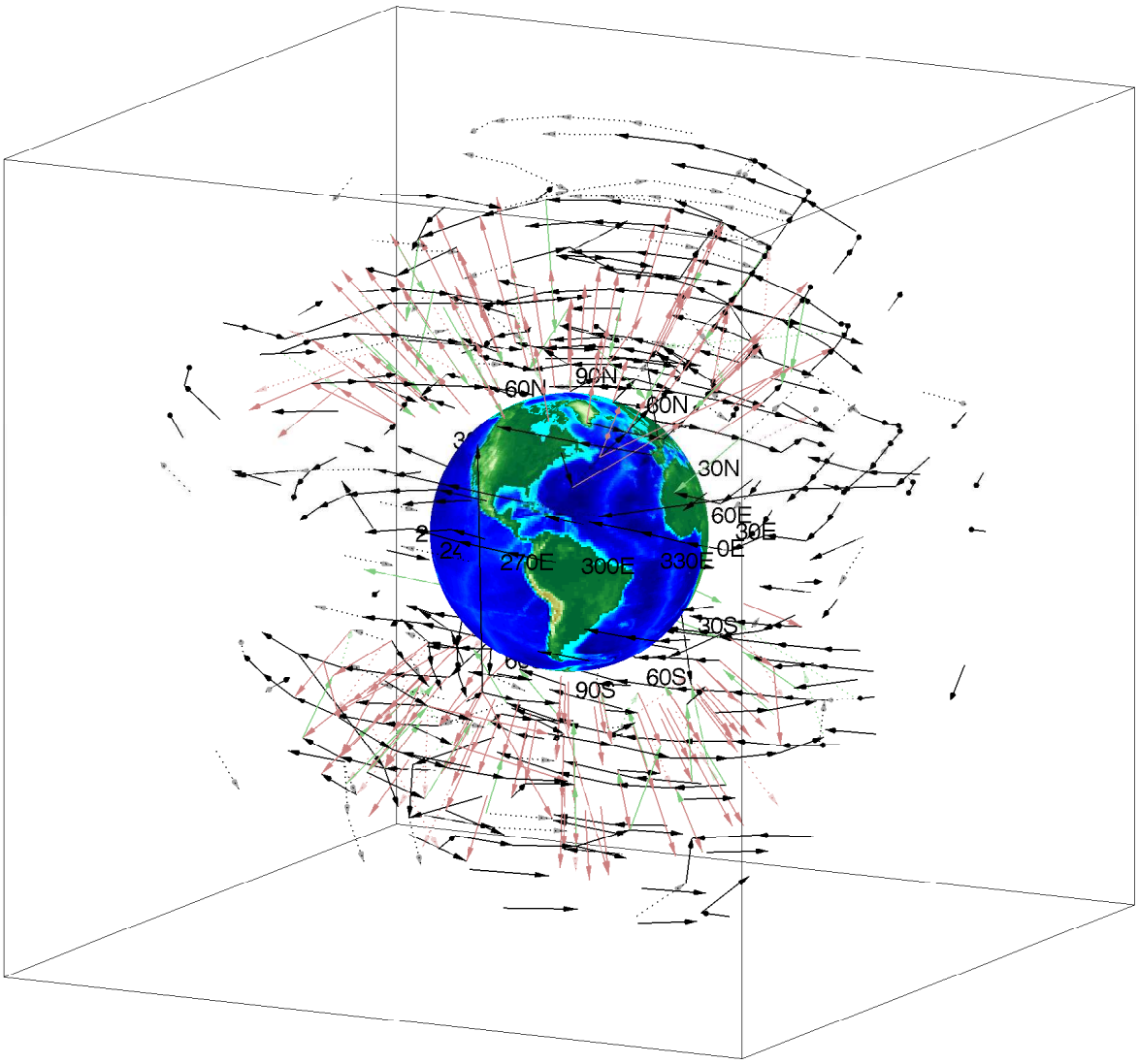


Figure 4: Spherical plot for 4 layers, strongest connections with travel time 1 day

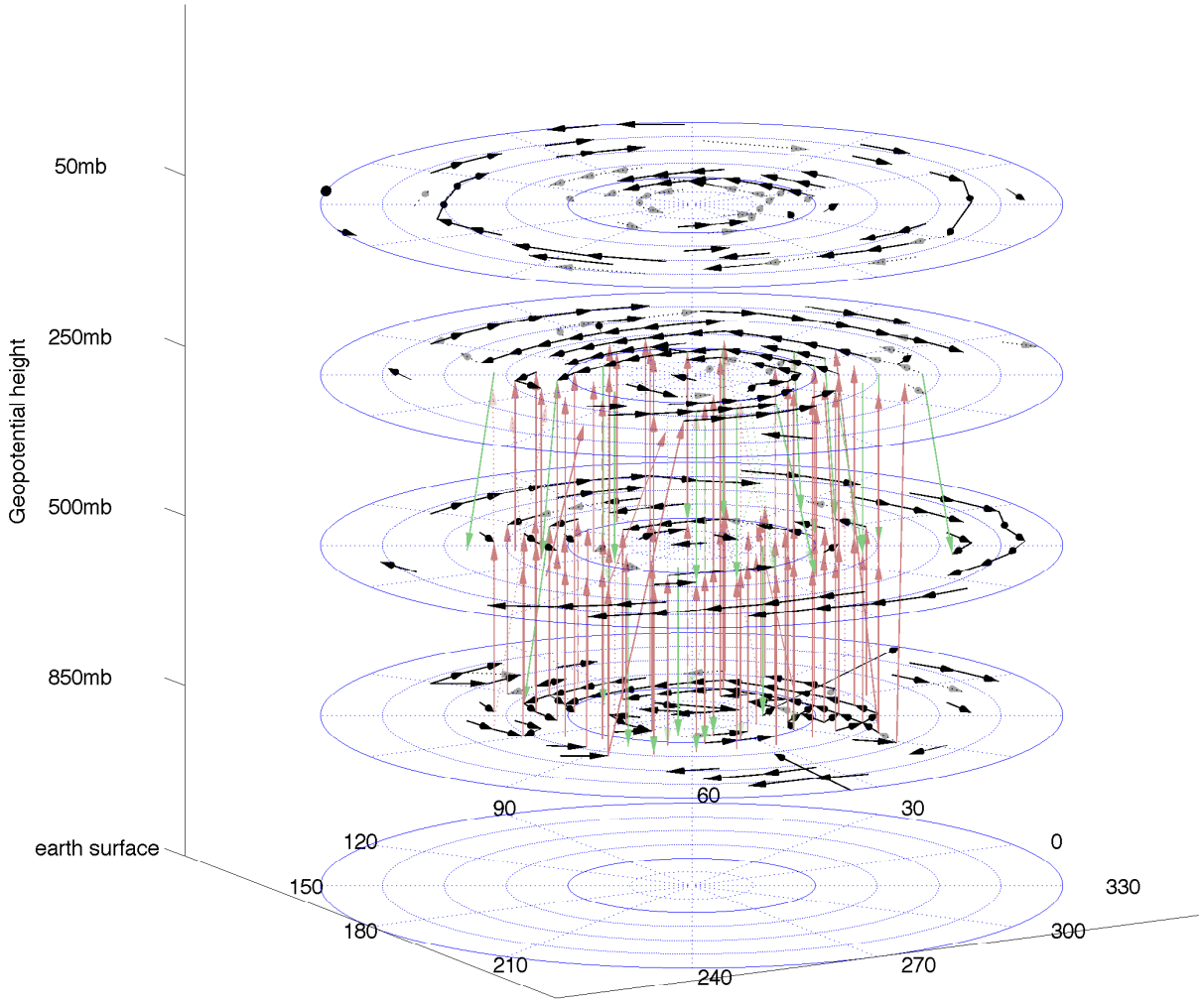


Figure 5: Stereographic projection plot for 4 layers, strongest connections with travel time 1 day, Northern hemisphere

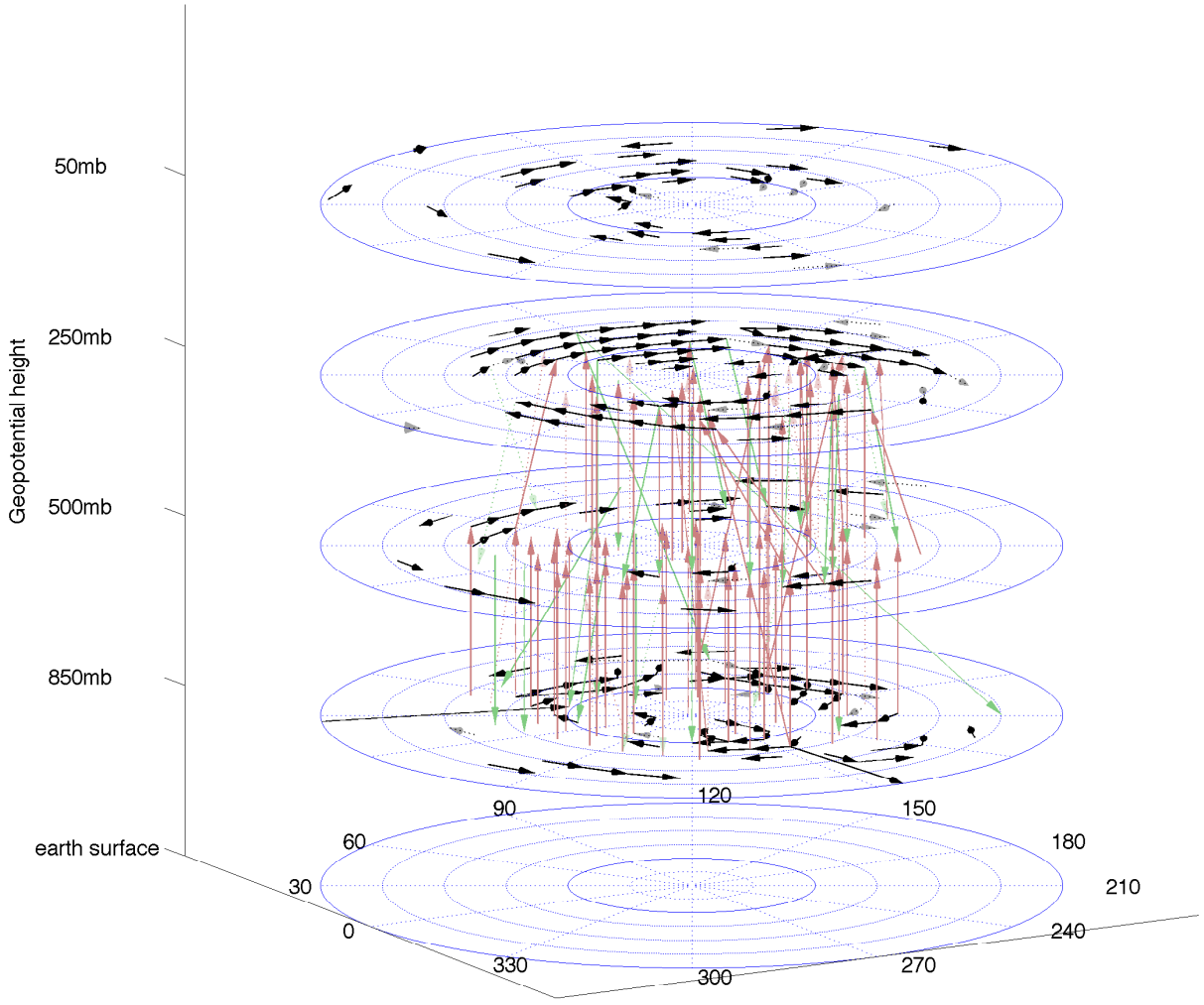


Figure 6: Stereographic projection plot for 4 layers, strongest connections with travel time 1 day, Southern hemisphere

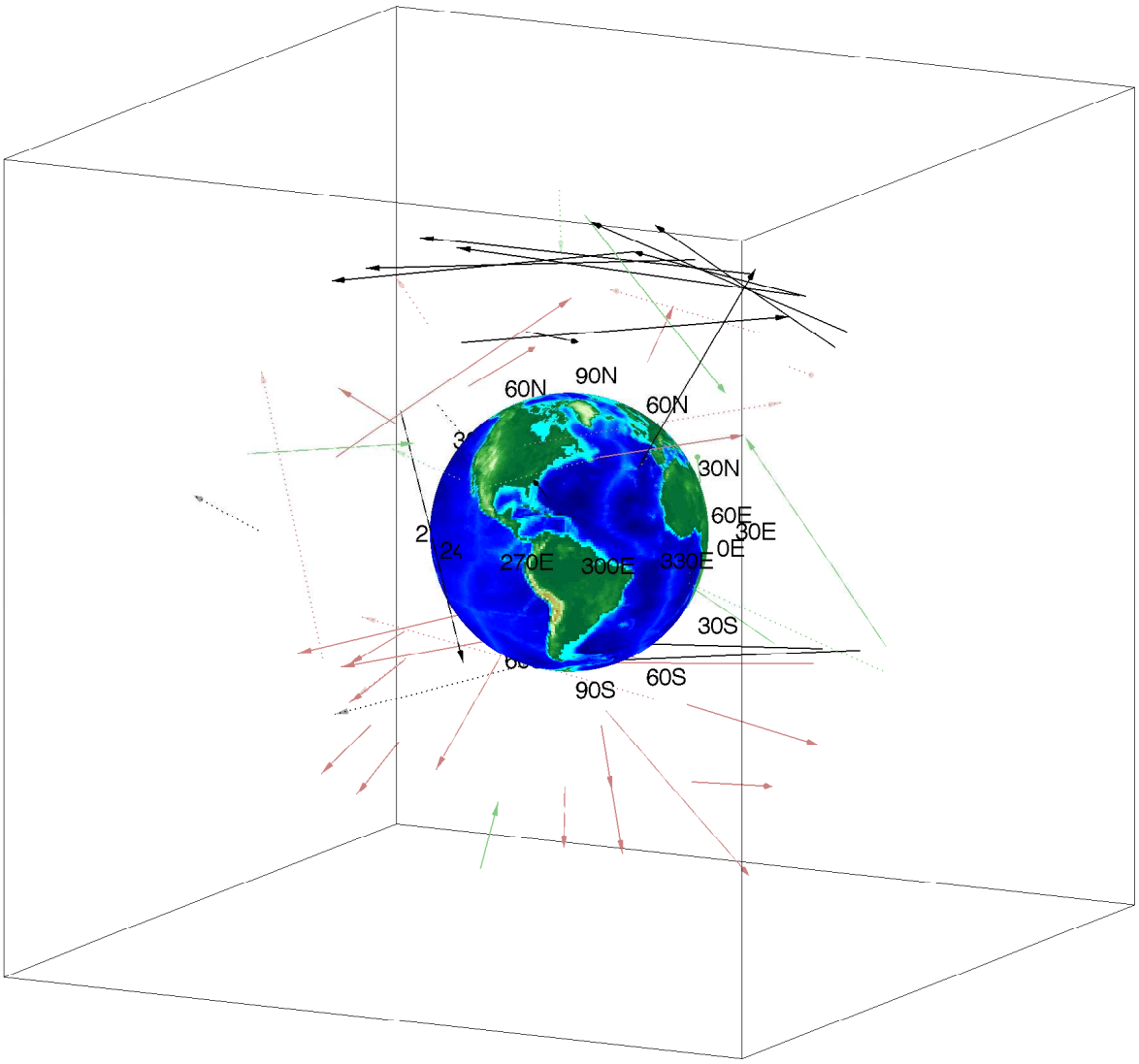


Figure 7: Spherical plot for 4 layers, strongest connections with travel time 2 days

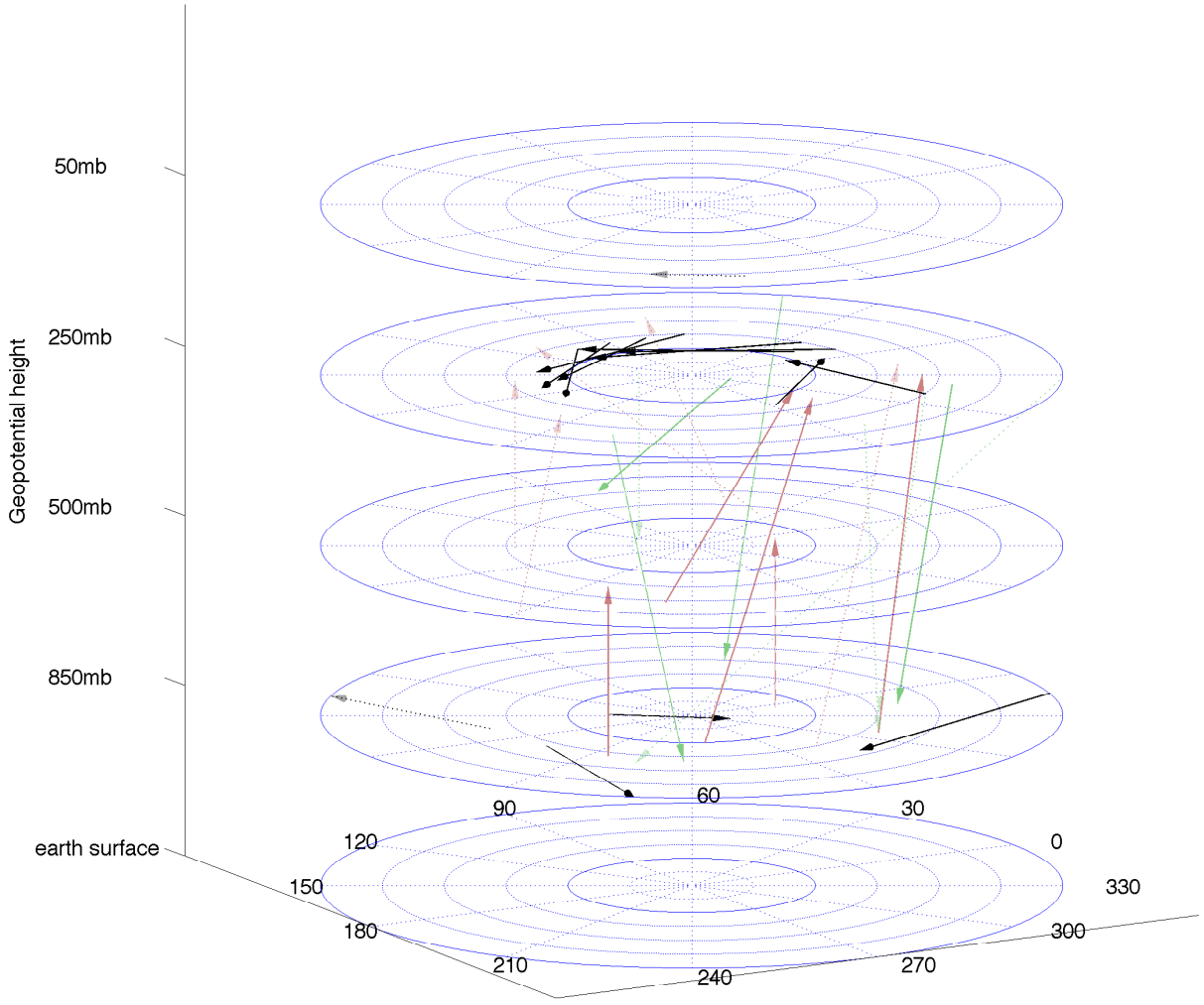


Figure 8: Stereographic projection plot for 4 layers, strongest connections with travel time 2 days, Northern hemisphere

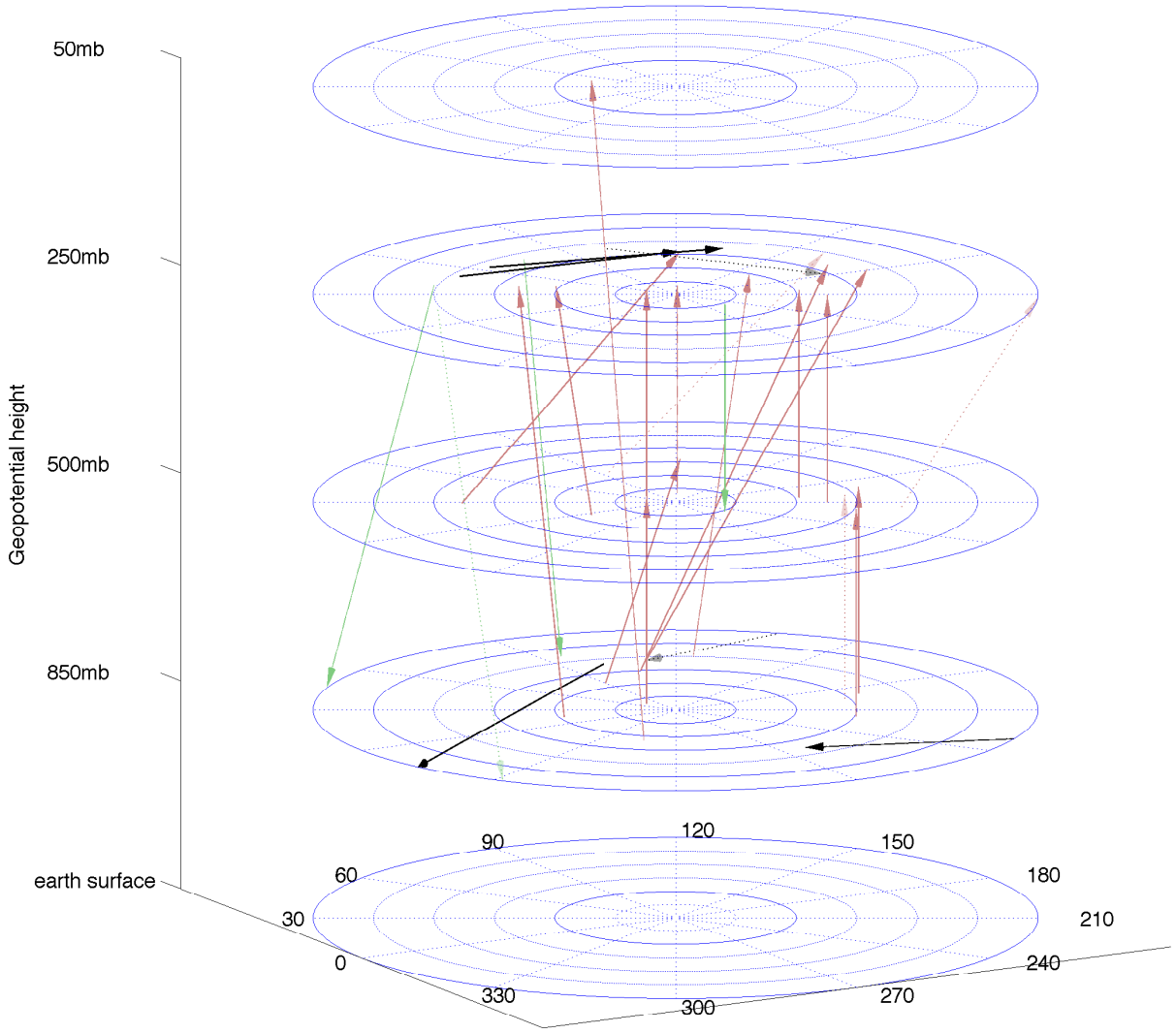


Figure 9: Stereographic projection plot for 4 layers, strongest connections with travel time 2 days, Southern hemisphere

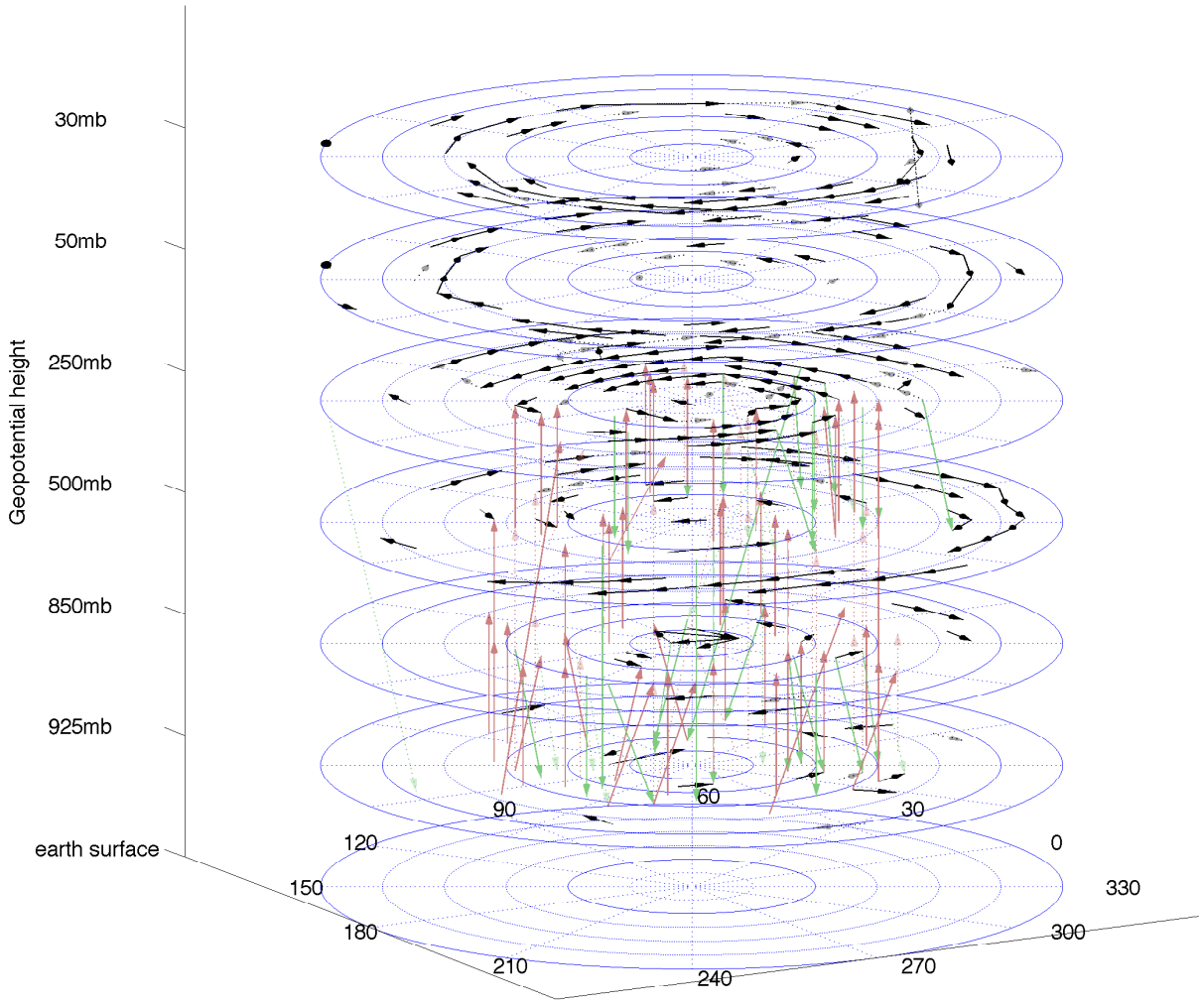


Figure 10: Stereographic projection plot for 6 layers, strongest connections with travel time 1 day, Northern hemisphere

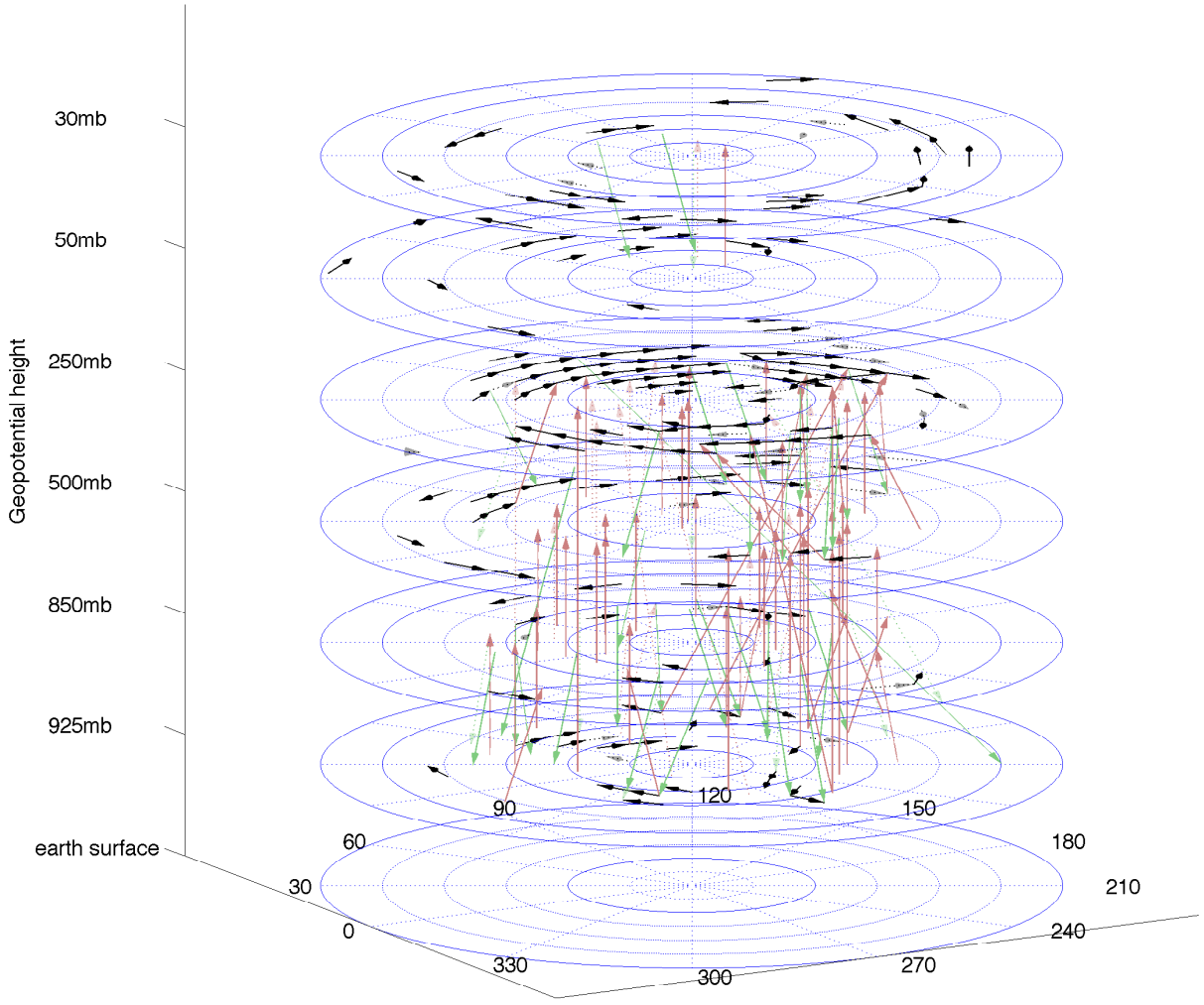


Figure 11: Stereographic projection plot for 6 layers, strongest connections with travel time 1 day, Southern hemisphere