A Dynamical History of the Inner Neptunian Satellites -and- Martian Weather: Viking Observations and M.O. Data Assimilation Techniques

Thesis by
Don Banfield

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1994
(Submitted December 10, 1993)

Acknowledgements

Academically, I would like to first acknowledge the help of my thesis advisor, Andy Ingersoll. He has been very patient, encouraging, and wise. I also appreciate his human nature and his quirky mannerisms. Thanks. I would also like to directly thank Norm Murray for advising me through the first chapter of this thesis. I admire him for his ability to be a brilliant researcher, yet also a regular guy. Again, thanks. I would also like to acknowledge all of the smaller chunks of help that all of the different faculty and fellow students have given me over the years here. It has been a very enlightening environment.

Looking farther back in time, I want to also acknowledge my family, who guided me to this great ziggurat of knowledge. I guess it was tearing apart old projectors and washing machines that got me interested in science. It probably helped to have curiosity re-enforced and encouraged. Mom, thanks. You're the best. I wish Dad could be here too. To my brothers and sisters, whatever... enough of this mushy crap.

To my surrogate family at Steuben house, I offer a fine $1 beer at the Colorado. If anything has been fun at Caltech, it has been because our house

had some life in it. Thank god I found you guys to live with, it has made all the difference. By my surrogate family at Steuben, I mean also all of the folks that have come in and out of that dive on a regular basis. That took guts, and I admire that as well. I'm sure we'll all look back on all of it someday and laugh. Of course, none of this really happened.

Abstract

We examine a scenario involving the capture origin of Triton, and infer the dynamical history of the Neptune satellite system. Triton's post-capture orbit forced chaotic perturbations on the original inner satellites of Neptune, leading to their mutual collisions and self-destruction. Neptune's current inner satellite system re-formed equatorially after Triton's orbital circularization. The 4.7° inclination of 1989N6 is probably due to a temporary inclination resonance. The 2:1 secondary resonance of the 1989N6-1989N3 12:10 resonance would eject 1989N6 at 4.7°, matching the observations. We have established limits for Neptune's Q: $12,000 < Q_N < 330,000$.

We examine a steady-state scheme for data assimilation in the context of a single, sun-synchronous, polar-orbiting satellite. The optimal (Wiener) gains are steady in time, and equivalent to those of a Kalman filter. The gains are computed by iteration using prior estimates to assimilate simulated observations of one model run ('Truth') into another run. The resulting prediction errors then form the next estimate of the gains. In model tests, the scheme works well even if only the mass field is observed. Although the scheme was developed for Mars Observer, it should be applicable to data

retrieved from Earth atmosphere satellites, e.g., UARS.

Spring and fall Viking IRTM T15 observations are used to estimate the Martian weather correlation length scale in the range 0.5-1 mbar. The results are important in providing a benchmark for validating Martian GCMs, determining the optimal placement of a network of landers, and guiding data assimilation efforts. Atmospheric temperature observations are used to compute an atmospheric mean state, which is subtracted from the observations to yield weather temperature residuals. These residuals are correlated with each other to determine the weather temperature correlation length scale ($\sim$ 1500km) and the weather temperature variance ($\sim 4-11K^2$). This work suggests that $\sim$ 110 landers are needed to globally observe Mars' weather.

Contents

Summary

This thesis consists of three papers. The first has already appeared as BANFIELD, D. AND N. MURRAY, 1992: "A dynamical history of the inner Neptunian satellites," *Icarus*, **99**, 390-401. It was the culmination of the work I did during my first 3 years at Caltech. I did a proposition with Peter Goldreich on the formation of Triton, which progressed into a paper titled "Neptune's Story." This paper outlined what is taken to be the most plausible explanation for the origin of Triton and the strange satellite system of Neptune. However, this was all done before the Voyager Neptune encounter. In that encounter, Voyager discovered 6 new satellites of Neptune, but they weren't quite oriented as predicted from the "Neptune's Story" paper. Thus, I examined what could be learned from the Voyager information about the satellites in light of the Triton formation story that we had published earlier. Among the conclusions were an explanation for the inclination of the smallest, innermost Neptunian satellite, and bounds on the Q value of Neptune.

As celestial mechanics was not the right field in which for me to excel professionally, I switched abruptly to Martian weather. In preparation for Mars Observer, Andy Ingersoll was attempting to put together a scheme for

assimilating Mars Observer PMIRR data into a global model of the Martian atmosphere. Andy and I learned much about data assimilation, and eventually came up with a technique that is nearly as optimal in its use of data as the mathematically optimal Kalman filter, yet numerically feasible to implement with today's computers. This steady-state Wiener filter is the subject of the second paper in this body of work. This chapter has been accepted by the Journal of Atmospheric Sciences. My co-authors are Andrew P. Ingersoll, and Christian L. Keppenne. Although the work was done when Mars Observer was a perfectly healthy spacecraft, and thus directed at using the imminent data from the spacecraft, it can be applied to almost any regularly orbiting spacecraft probing an atmosphere.

The third chapter is directed at investigating properties of Martian weather that we can already see in existing data sets. Specifically, it is a study of the weather correlation length scales on Mars, using Viking IRTM observations. My co-authors are Andrew P. Ingersoll, Anthony D. Toigo, and David A. Paige. These investigations are needed to specify the first-cut gain functions for a data assimilation system such as the one discussed in the second chapter. They also can be used to suggest optimal placement densities of meteorological landers, such as the MESUR network proposed to be sent to Mars in the next decade.

As planetary science is a risky business, we go through great highs and serious lows. In the first part of my Caltech career, I worked on the imaging team of the Voyager Neptune encounter, a unique and extremely exciting experience. Then I started preparing for another mission, Mars Observer. After a few years of work, I (probably less than many others) was extremely

disappointed to see the mission fail. Nevertheless, we look ahead to what we can do in the future, and re-direct our work towards the questions that we still can and want to answer.

Chapter 1

A Dynamical History of the Inner Neptunian Satellites

1.1 Abstract

We examine a scenario involving the capture origin of Triton, and infer the probable dynamical history of the Neptune satellite system. Triton's highly elongated post-capture orbit forced chaotic perturbations of the eccentricities of the original inner satellites of Neptune. The resulting large eccentricities quickly led to mutual collisions and self-destruction of the inner satellites, leaving a disk of debris. Thus, Neptune's inner satellite system re-formed on equatorial orbits after Triton's orbital circularization. It is unlikely that these re-formed satellites survived the flux of cometary impactors. Hence, today's satellites are probably the remnants of parent satellites destroyed even more recently than Triton's capture.

The remaining puzzle is the inclination of 1989N6, at 4.7°. This is probably due to a temporary inclination resonance capture. 1989N6 has passed through ≤ 35 inclination resonances in its history. The probability of avoiding all of these is ~24%; we suggest that 1989N6 was temporarily caught in one. The 1989N6-1989N3 12:10 resonance has a 2:1 secondary resonance that would eject 1989N6 from the primary resonance at an inclination of roughly 4.7°, matching the observations. We have established the following limits for the Q value of Neptune: $12,000 < Q_N < 330,000$. The upper limit is found by requiring the resonant configuration to have occurred during the satellite's lifetimes, and the lower limit is found from the requirement that 1989N1 not tidally evolve through corotation.

1.2 Introduction

Three of the giant planets possess regular satellite systems; Neptune does not. Instead, it has one massive satellite, Triton, on a highly inclined, retrograde orbit, a small distant satellite, Nereid on a very eccentric and inclined orbit, and a regular set of small inner satellites. The smallest, and innermost of this regular set is also inclined by $\sim$ 4.7°. Thus, the origin and dynamical history of this system is clearly different than that of the satellite systems of the other three giant planets.

A number of workers (Farinella *et al.* 1980; McKinnon 1984; Goldreich *et al.* 1989) have suggested that Triton was once a rogue body in Heliocentric orbit, which was captured by Neptune. This hypothesis can explain many of the features of Neptune's satellite system. But it only partly solves

the puzzles surrounding the small inner satellites. This paper addresses the dynamical history of these inner satellites.

This work is based upon the scenario outlined in Goldreich *et al.* (1989) (hereafter referred to as GMLB), in which they proposed that Triton was captured from a heliocentric orbit by collision with a primordial satellite of Neptune. Triton's orbit then circularized due to tides. In examining this proposition in detail, GMLB made predictions about what would be observed in the Neptune system by the Voyager spacecraft. We re-examine their conclusions in light of the Voyager encounter with Neptune in 1989, extending the scenario to explain some of the new discoveries made during the encounter.

Before Voyager, the only known constituents of Neptune's satellite system were Triton, Nereid, and what were believed to be ring arcs. Triton was known to be a large satellite on a circular retrograde inclined orbit. Nereid, a small body, was observed to follow a large, prograde inclined and eccentric orbit about Neptune. The ring arcs were found close to Neptune, at only a few planetary radii. GMLB concluded that the most plausible scenario for generating this system started with the capture of Triton from heliocentric orbit. Subsequently, tidal dissipation reduced the initially large eccentricity and semimajor axis of Triton's orbit, leaving the inclination unchanged.

Early in this tidal evolution, Triton's orbit passed through the region in which Nereid orbits Neptune. GMLB showed that the perturbations caused by Triton on Nereid's orbit were sufficient to account for Nereid's irregular (both inclined and eccentric) orbit. They also noted that Triton passed through Neptune's equatorial plane numerous times. They inferred that this

would have caused the destruction of any primordial satellites outside of 5 Neptune radii (R_N). Voyager observations confirmed this: six new satellites were discovered, all inside 4.7 R_N. Thus, the capture scenario of GMLB successfully predicted the nature of Neptune's inner satellite system.

Nevertheless, there were predictions made by GMLB that were not found to be accurate. They predicted that there would be a satellite inside of 5 R_N on an inclined orbit. This satellite was supposed to be large enough to act as a shepherd to the ring arcs. The satellite was predicted to be on an inclined orbit, due to chaotic perturbations forced by Triton. Triton would have imparted a large eccentricity to the satellite as well, but GMLB noted that any such eccentricity would be damped out by tides on a time scale short compared to the age of the solar system.

The Voyager observations of the inner Neptunian satellite system revealed not one inclined moon, but six moons, five of which are on nearly equatorial orbits. The sixth, the innermost and smallest satellite discovered, follows an orbit inclined by about 4.7°, and is too small to act as a shepherd to the ring arcs (Smith *et al.* 1989). Thus, it appears that there is more to the history of the Neptune satellite system than was conjectured in GMLB.

In the second section of this paper, we examine the implications of the above scenario for the early history of the original inner satellites of Neptune. The third section considers the effects of tidal evolution on the satellites. We also derive limits on the tidal quality factor, Q, of Neptune. The fourth section explains the inclination of 1989N6 in terms of a temporary resonance capture. We then discuss the effects of similar resonances on the other inner satellites, followed by our conclusions.

1.3 Self Destruction of Original System

The conclusions of GMLB were based on the implicit assumption that the interior satellites of Neptune did not collide with one another. As mentioned above, GMLB showed that perturbations due to Triton during the time its orbit circularized would have caused increases in the eccentricity and inclination of the orbits of inner Neptunian satellites. However, if there were multiple inner satellites, the possibility arises that they would have been perturbed onto crossing orbits and hence destroyed.

The eccentricities imparted to a satellite by Triton during its orbital circularization can be estimated as a function of the satellite's semimajor axis. GMLB stated that the Triton perturbations can be expressed as a function of the semimajor axis of the satellite. For eccentricity, this function is $\Delta e \propto (a/5R_N)^5/\sqrt{1-a/5R_N}$. GMLB also showed that the satellites' evolution would be chaotic, and thus can be modeled as a diffusive process, with eccentricity increasing as the square root of time. We found the constant of proportionality in this relation by fitting it to the curves in Fig. 6 of GMLB, which shows the results of a 7×10^7 year calculation of the eccentricity of test satellites. GMLB noted that when the satellites' eccentricities are large, they are also subject to damping under satellite tides. Therefore, in our model of the effects, we limit the eccentricity gained during the Triton perturbations to about 0.3, above which tides are more effective at reducing the eccentricity than the Triton perturbations are at raising it. The value of 0.3 is consistent with the results in Fig. 6 of GMLB. This analysis yields a function which is near zero at $2R_N$ and quickly rises to $e \sim 0.3$ for $a \rightarrow 5R_N$. With these

eccentricities, one satellite just outside of the Roche limit (at $\sim 2.7R_N$ for ice) would collide with another just inside $5R_N$. If there were more satellites in this region, as there are now, the problem would only be worsened. Therefore, we believe that any reasonable set of parent satellites in orbit around Neptune at the time of Triton's capture would have been on crossing orbits for some period of time. The timescale for which these orbits overlap is the eccentricity damping time for these small satellites, roughly 10^8 years.

We examine the likelihood of mutual collisions between the inner satellites. The orbital periods of these satellites are of order half a day, and their precession times are of order 3 years. For times much greater than 3 years, the longitudes, lines of nodes and lines of apses may be considered random. Without loss of generality, we take a satellite (satellite 1) to be in the equator plane, and another, satellite 2 to be inclined to that. We denote the satellite radius, semi-major axis and eccentricity by r, a and e. Our attention is then focussed on the time when the equatorial satellite is on the other's line of nodes. We assume that satellite 1 can be found anywhere along this line from apocenter to pericenter, with equal probability. Satellite 2 is somewhere along its orbit, but the probability that it is a satellite radius or less from the line of nodes is just $r_2/\pi a_2$. The next requirement for collision to occur is that the first satellite lies between the second's apoapse and periapse. This probability is given by $[a_2(1+e_2)-a_1(1-e_1)]/[a_1(1+e_1)-a_1(1-e_1)]$ where we have assumed that the satellite 1 is the interior one and that the eccentricities are large enough that the denominator is greater than a satellite radius. Finally, we require that the second satellite be within the satellites' radii of the first satellite, along the line of nodes. This probability is just

$(r_1 + r_2)/a_2 e_2$. Thus, the probability of a collision per orbit of the first satellite is the product of these three quantities. We take $a_1 \sim a_2 \sim 3.5 R_N$ and $e_1 \sim e_2 \sim 0.3$ and $r_1 \sim r_2 \sim 80$km which allows us to simplify the equation to:

$$\frac{P_{col}}{orbit} \sim \frac{2r^2}{\pi a^2 e}. \tag{1.1}$$

This yields a collision timescale of about 10^3 years. Thus, mutual collisions would almost certainly occur over the 10^8 years it would take the eccentricities to damp out. Note that this calculation is roughly equivalent to calculating the number of satellite orbits required for a satellite's cross-section to fill in the area in the equator plane accessible to it.

A collision between any of Neptune's current inner satellites and its neighbor would easily destroy both of the satellites involved. This can be shown with the following relation from Stevenson *et al.* (1986):

$$\frac{1}{2} M_{body} V_{impact}^2 \sim M_{sat} S + \frac{3}{5} \frac{G M_{sat}^2}{\gamma r_{sat}} \tag{1.2}$$

with $S \sim 10^6 \text{erg} \cdot \text{g}^{-1}$ and $\gamma \sim 0.1$. This says that the energy of the impacting body must be roughly equal to the energy, S, stored in material strength of the target body plus the gravitational energy of the target body in order to break it up. For the eccentricities expected, about 0.3, all but 1989N1 could be easily destroyed by an impact with 1989N6, the smallest satellite, and 1989N1 could be destroyed by 1989N2, 1989N3 or 1989N4. We conclude that the current satellites could not have existed prior to the time of Triton's orbit circularization.

We propose that the original inner satellite system of Neptune was like the present inner satellite system, but was destroyed through mutual collisions. The shards of the collisions quickly damped out all remaining eccentricities and inclinations due to further collisions, and thus erased any signature of Triton's chaotic perturbations. After Triton's orbit circularized, new satellites formed out of this equatorial disk of rubble within $5R_N$, the remnants of the primordial inner Neptunian system. The daughter satellites had no "memory" of the inclination or eccentricity perturbations given to their parent bodies by Triton.

To back up these assertions, we examine the magnitudes of the timescales and processes involved. The timescale for a rubble disk to damp to the equator plane can be estimated by looking at the optical depth of such a disk. We model the remnants of the original system as a distribution of bodies following a power-law size distribution, $n(r)dr = n_o r^{-3} dr$ cm^{-3}, (e.g., Burns *et al.* 1984, p. 211). We choose a lower limit of $r_l \sim$ 1cm, the size at which sticking forces are significant in the accretion of bodies (Longaretti 1989, p. 54). For the upper bound we choose $r_u \sim$ 8km (Harris 1984, p. 653). Neither assumption strongly affects the results. We further assume that the total mass of 1989N1-1989N6, M_{sats} is spread out evenly in a disk of height h over the range of semimajor axes, $a_i = 1.9R_N$ to $a_o = 4.7R_N$.

$$M_{sats} = \frac{4}{3}\pi^2 \rho n_o (a_o^2 - a_i^2) h (r_u - r_l) \tag{1.3}$$

We then wish to determine the optical depth of this disk.

$$\tau = hn_o\pi \ln \frac{r_u}{r_l} \tag{1.4}$$

Solving (1.3) for hn_o and employing the result in (1.4) we find $\tau \sim 10^{-3}$. We can then estimate the inclination (or eccentricity) damping time scale by multiplying the orbital period of the particles by the inverse of the optical depth. This calculation yields timescales on the order of 1 year. Therefore, it is safe to say that an inclined and eccentric rubble disk will quickly damp to an equatorial disk.

The timescale for re-forming the satellites is an important, although more difficult question to address. We estimate the accretion time using a simple formula from Stevenson *et al.* (1986), assuming a disk of material as above.

$$\tau_{acc} \sim \frac{m}{\pi r^2 \sigma_s \Omega} \tag{1.5}$$

where m is the mass of one of the satellites, r is its radius, Ω its orbital frequency and σ_s is the surface density of the proto-satellite disk. This yields a timescale of about 10 years. Because the satellites are near (or within) the Roche limit, the accretion would probably not be as efficient as is assumed in this calculation. Nevertheless, the re-accretion timescale is quite short compared to the age of the solar system. Thus, the rubble quickly settles down to a flat disk, then slightly more slowly re-forms satellites.

It is unlikely that the sizes or orbital configuration of the primordial satellites were passed on to their daughter bodies. This is because, given the

eccentricities involved (~ 0.3), the collision velocities were of the same order as the orbital velocities, and we can presume that the masses of the colliding bodies (the primordial satellites) were also of the same order. Therefore, energy exchanges in the collisions were large enough to have caused significant semi-major axis changes, $\Delta a/a \sim \mathcal{O}(1)$ in addition to destroying the bodies involved. We expect the re-accreted bodies to bear little resemblance to their parents.

1.4 Orbital Evolution

To track the history of the satellites from the time of Triton's orbital circularization to the present, we must recognize that other effects have had much influence on the system. Smith *et al.* (1989), and Colwell and Esposito (1992), estimate the flux of cometary bodies passing through the Neptune system in this time period. They show that this flux of impactors would have destroyed all of the current inner satellites except 1989N1 over the lifetime of the solar system. These authors derive expected lifetimes for the current satellites, with that for 1989N6 being as low as 1 to 2 billion years. The result of a collision between a substantial comet and one of the inner satellites would be similar to the evolution following mutual collision disruption: damping of the rubble from the parent satellite to a disk, possibly followed by equatorial re-accretion.

Tides raised by the satellites on Neptune have also had significant influence on the system over this time period. Furthermore, by examining the tidal evolution of the inner satellites, we can derive upper and lower limits on

the Q value of Neptune. Because 1989N1 is the only satellite likely to have survived intact since the time Triton's orbit circularized, T_{circ}, it is the only one we can use to set a lower limit on Q_N. Tides raised on Neptune by a satellite affect that satellite's semimajor axis (e.g., Goldreich and Soter 1966). Prograde satellites outside corotation move away from the planet, while satellites inside corotation move inward. At the semimajor axis corresponding to corotation with the planet, there is no evolution of the satellite's semimajor axis. Therefore, because we observe 1989N1 outside of corotation, it was never inside that point. The lowest value of Q_N that would place 1989N1 at corotation just after Triton's orbit circularized is $12,000 \times \frac{4 \times 10^9 \text{years}}{T_{circ}}$. Note, however, that this and other calculations in this paper assume a time and frequency independent value for Q_N.

We can also derive an upper limit for Q_N because, as we will discuss below, 1989N6 must have gone through a particular inclination resonance to explain its currently observed 4.7° inclination. That is, there must have been enough tidal change in the satellites' orbits for them to have once been in this resonant configuration. Furthermore, this had to have occurred since the time 1989N6 was created by the disruption of its parent body. Because this time is not well known, we will derive the upper limit under the weaker constraint that this occurred since Triton's orbit circularized. This yields a value of $Q_N \leq 330,000 \times \frac{4 \times 10^9 \text{years}}{T_{circ}}$; probable limits are one half to a quarter of this value since the survival time for 1989N6 against cometary impacts is 1 to 2 billion years.

As Fig. 1.1 shows, the four innermost moons are inside the Roche limit ($\sim 2.7 R_N$ for icy satellites). However, this situation is not unique in the Solar

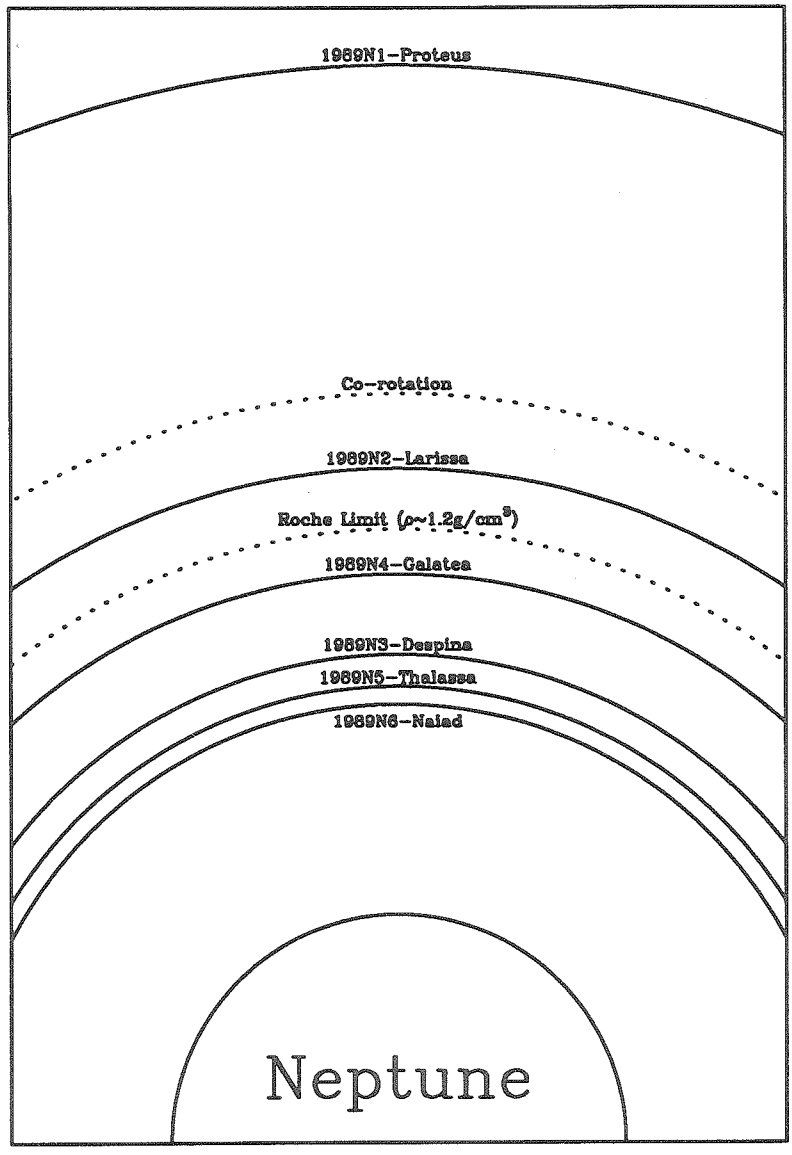


Fig. 1.1: This shows the inner Neptunian satellite system in plan view at the current time. The planet is at the bottom, the orbits of the satellites are shown to scale, and the rings have been left out for clarity. We have also indicated the location of the corotation orbit and the Roche limit for icy bodies. Note that there are four satellites currently inside the Roche limit.

System. Jupiter has two satellites well inside the Roche limit and Uranus has five. The average radius of these bodies is $\sim$ 20km. Accretion is possible within the Roche limit, as non-gravitational forces can allow bodies to stick together after collisions. This is particularly true for bodies of this size and smaller, where material strength is a force comparable in magnitude to that of gravity. However, accretion is not the only mechanism for producing satellites inside the Roche limit. We sketch here a possible scenario for the satellites' history which could yield the current distribution. We do not suggest that this scenario is unique, merely that it is possible. 1989N1 has probably survived since the time of Triton's orbital circularization, moving out to its present location under the action of planetary tides. 1989N2-1989N4 now straddle the Roche limit, but could easily have all formed outside of it, then migrated inwards due to planetary tides. It is also possible that 1989N5 and 1989N6 are the remnants of an impact on a larger parent body. This parent body could have formed beyond $2.7R_N$ and migrated inside the Roche limit due to the stronger tides produced by its larger mass. We conclude that the satellites could have accreted within the Roche limit or arrived there under the action of tides and collisions.

1.5 Inclination of 1989N6

The only puzzle that the above scenario leaves about the inner satellites is that 1989N6 is elevated out of the equatorial plane by 4.7°. The five other newly discovered satellites all have orbits that lie very near to the equator plane of Neptune, as we would expect from the formation scenarios sketched

above. What produced the inclination of 1989N6?

One possibility is that a collision between 1989N6 and a passing comet explains the satellite's observed inclination of 4.7°. However, this comet must not destroy 1989N6; otherwise it would just re-accrete from an equatorial disk of debris. Using the collisional breakup calculation from above, (1.2), and a cometary impact speed of $\sim 14\text{km} \cdot \text{s}^{-1}$, we find that the largest comet that could hit 1989N6 without destroying it is only $\sim 10^{-5}$ the mass of 1989N6. Such an impact would impart an inclination of roughly that same magnitude, well under the 4.7° that is observed. We conclude that an impact by a comet cannot produce the observed inclination of 1989N6.

Uranus' satellite, Miranda is observed to be inclined by a few degrees, much like 1989N6. A possible explanation for this inclination is given in Tittemore and Wisdom (1989) (T&W89), and in Malhotra and Dermott (1990) (M&D90). These authors demonstrate that an inclination resonance between Miranda and one of the other satellites of Uranus probably existed for some time in the past. A consequence of this occurrence is that Miranda's inclination grew as the satellites' orbits evolved under tidal dissipation in the resonance. Eventually, Miranda was ejected from the resonant configuration due to a secondary resonance, leaving Miranda with its current inclination. 1989N6 could have experienced a temporary resonance capture in a manner much like Miranda. Therefore, in the remainder of this paper we will explore the possible resonant states that the inner Neptunian satellite system passed through since its formation, with a watchful eye for those resonances that could have resulted in the observed inclination of 1989N6. The responsible primary-secondary resonance combination should not only have a reasonably

Table I: Inner Neptunian Satellites Physical Characteristics[1]

Name		Radius	a	i	e	Mass[2]
1989N1	Proteus	208 ± 8km	$4.75R_N$	0.04°	0.0004	4.0×10^{22}g
1929N2	Larissa	96 ± 7km	$2.97R_N$	0.20°	0.0014	4.3×10^{21}g
1989N4	Galatea	79 ± 12km	$2.50R_N$	0.05°	0.0001	3.7×10^{21}g
1989N3	Despina	74 ± 10km	$2.12R_N$	0.07°	0.0001	2.1×10^{21}g
1989N5	Thalassa	40 ± 8km	$2.02R_N$	0.21°	0.0002	3.2×10^{20}g
1989N6	Naiad	29 ± 6km	$1.94R_N$	4.74°	0.0003	9.9×10^{19}g

[1] The data for this table was taken from Owen *et al.* (1991) and Thomas and Veverka (1991).

[2] The masses are calculated assuming $\rho \sim 1.2$ g·cm^{-3}

likely capture probability for the primary resonance, but also that primary resonance must be stable against tidal forcing. Furthermore, the secondary resonance involved should also have a reasonable capture probability and be stable against tidal forcing as well. The secondary resonances must also cause the satellites to be ejected from the primary resonance at an inclination roughly matching 1989N6's current 4.7°.

In order to identify possible resonance passages, Fig. 1.2 shows a model of the tidal evolution of the inner Neptunian satellites, with Q_N chosen to start 1989N2 just inside corotation at time, T_{circ}. Had there been no cometary impacts, this would be the most possible tidal evolution the system could have experienced. The values of the masses used are given in Table I. The masses are very poorly constrained for these satellites, so the calculations were also performed for reasonable limiting cases. However, the curves are not significantly affected by this. We have restricted our attention to second-order resonances because those are the lowest order resonances that involve, and thus can affect, the inclination. The vertical bars on Fig. 1.2 represent

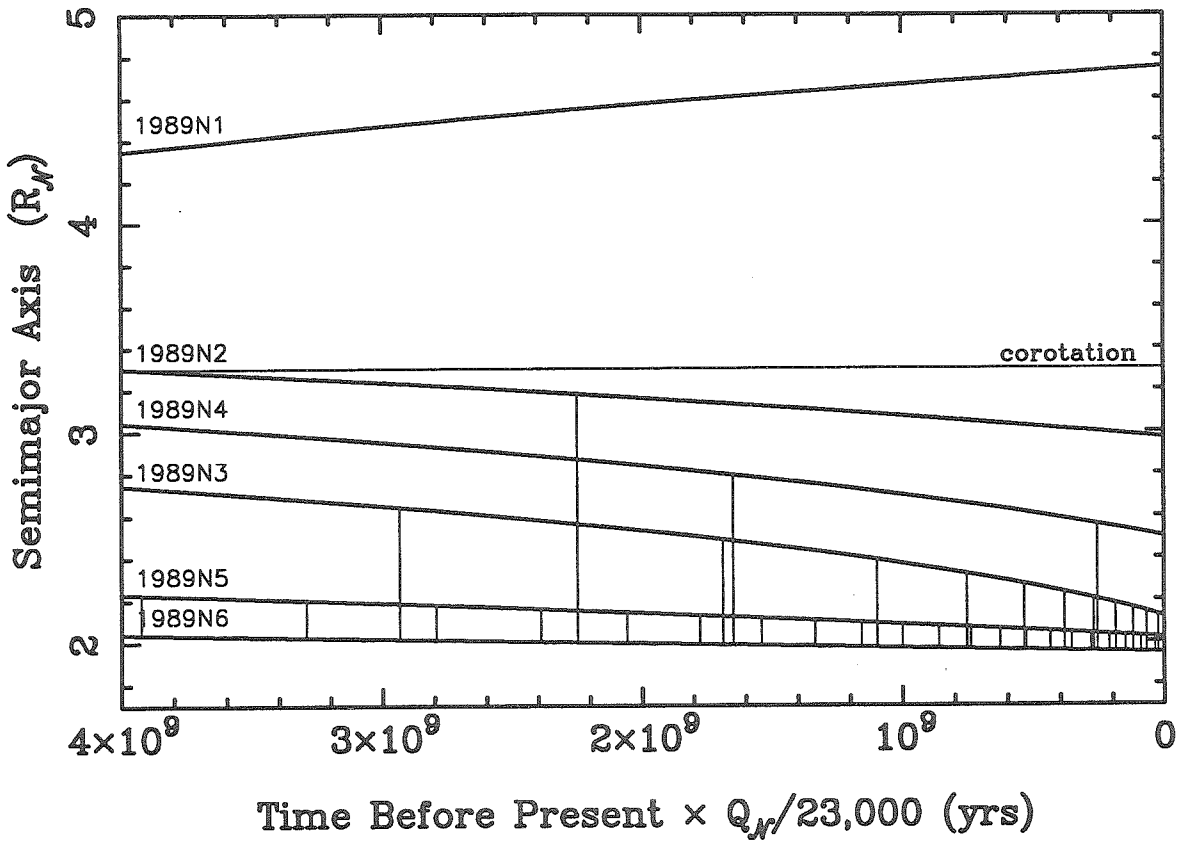


Fig. 1.2: A model of the tidal history of satellites 1989N1-1989N6, calculated with Q_N assumed to be 23,000. The vertical lines represent times when 1989N6 was in a second-order commensurability with one of the other satellites. Note that none of the satellites cross corotation.

the times when the two satellites connected by a bar passed through one of these commensurabilities. Because we are interested in explaining a change in 1989N6's inclination, we have only indicated those involving 1989N6. Commensurabilities between 1989N6 and 1989N1 were also ignored, due to the fact that any resonance passages between these two satellites could not result in capture (e.g., Peale 1986, p. 178). 1989N1 is evolving outwards, and 1989N6 is falling in; so the divergence of their orbits causes possible resonances between them to have zero probability of capture. The other satellites are all falling towards Neptune faster than 1989N6 (a result of 1989N6's very small size, which outweighs its proximity to Neptune), and thus resonance passages with them could result in capture.

There are roughly 35 second-order mean motion commensurabilities that 1989N6 could have passed through since Triton's orbit circularization. They range from a 4:2 commensurability with 1989N2, to a 36:34 commensurability with 1989N5. Each of these mean motion commensurabilities actually consists of six resonances, three involving inclinations and three involving eccentricities. The names and arguments of these resonances are given in the order they are encountered by:

i_6^2 resonance	$\phi_{i_6^2} = (p+2)\lambda_x - p\lambda_6 - 2\Omega_6$
$i_6 i_x$ resonance	$\phi_{i_6 i_x} = (p+2)\lambda_x - p\lambda_6 - \Omega_6 - \Omega_x$
i_x^2 resonance	$\phi_{i_x^2} = (p+2)\lambda_x - p\lambda_6 - 2\Omega_x$
e_x^2 resonance	$\phi_{e_x^2} = (p+2)\lambda_x - p\lambda_6 - 2\varpi_x$
$e_6 e_x$ resonance	$\phi_{e_6 e_x} = (p+2)\lambda_x - p\lambda_6 - \varpi_6 - \varpi_x$
e_6^2 resonance	$\phi_{e_6^2} = (p+2)\lambda_x - p\lambda_6 - 2\varpi_6$

where the subscript x denotes the other satellite involved in the resonance. We have ignored the eccentricity resonances ($e_x^2, e_6 e_x, e_6^2$) for two reasons. First, we are interested only in the possible inclination effects of resonance capture, and eccentricity resonances do not affect inclination. Secondly, their perturbations on the other resonances are small. The magnitude of the perturbations between two resonances can be expressed by the ratio of the resonances' libration widths to their separations (Dermott *et al.* 1988). We have examined this ratio for all of the eccentricity resonances, both first and second-order, and for all of the inclination resonances. We found that, for reasonable eccentricities (e $\sim$ 0.0004) and initial inclinations (i $\sim$ 0.06°), the perturbations due to the eccentricity resonances are smaller than the perturbations due to the nearest inclination resonance. Therefore, it is safe to ignore the eccentricity resonances.

We have further restricted our attention to only one of the three inclination resonances, that which affects only the inclination of 1989N6, or the i_6^2 resonance. The inclination resonances are well separated from each other at the presumed initial inclinations and thus can be considered individually during capture. We ignore the third inclination resonance encountered, i_x^2, because it only involves the inclination of the other satellite in the resonance, and thus could not have increased 1989N6's inclination. The second inclination resonance, $i_6 i_x$, involves both inclinations, but has the effect of increasing both satellites' inclinations if they are captured into it. In order to increase 1989N6's inclination to 4.7°, the inclination of any of the other satellites would have had to be increased to a value higher than what we observe today. An unlikely possibility is that this did occur, but that a later cometary

impact then obliterated the remnant inclination of the other satellite in the resonance, thus reconciling this possibility with the observations. We believe it is safe to take this as so unlikely that none of the $i_6 i_x$ resonances were responsible either, and we are left with only the 1989N6 inclination resonance at each commensurability as a viable candidate.

We are now interested in the likelihood that the satellites will become captured into these i_6^2 resonances. Particularly good discussions of the mathematics involved can be found in Borderies and Goldreich (1984), Tittemore and Wisdom (1988), and Dermott *et al.* (1988), and we refer the reader there for the details of the calculations. A property of these resonances is that if the resonance is encountered with the satellite having an inclination below a certain value, then capture into the resonance is certain. The probability of capture falls off if encounter occurs with an inclination above this value. Table II lists the critical inclinations, and the resulting capture probabilities for the 35 possible resonance encounters. In calculating these quantities, we have assumed an initial inclination of 0.06° for 1989N6. This value is consistent with the current distribution of the inclinations of the other inner Neptunian satellites. Under this assumption, the resonances are all encountered at an inclination well above the critical inclinations. It is interesting to note that all of the resonances have capture probabilities between 2% and 10%. We can conclude from this that if the satellites passed through most of the commensurabilities, the chances that 1989N6 was caught into one of these is about 76%. Therefore, it is very likely that 1989N6 *was* temporarily caught into one of these resonances.

Are the resonances strong enough to actually retain the satellites' tra-

Table II: 1989N6 Primary Resonance Results

Resonance Sats	$p+2:p$	i_{crit}	$P_{capture}$	Q_{min}
N6-N5	16:14	.00138°	2.1	130
N6-N5	17:15	.00138°	2.1	110
N6-N3	6:4	.00431°	6.5	1500
N6-N5	18:16	.00138°	2.1	96
N6-N5	19:17	.00137°	2.1	85
N6-N2	4:2	.00677°	10.2	3000
N6-N5	20:18	.00137°	2.1	75
N6-N5	21:19	.00137°	2.1	67
N6-N3	7:5	.00422°	6.3	1200
N6-N4	5:3	.00596°	8.9	2300
N6-N5	22:20	.00137°	2.1	59
N6-N5	23:21	.00137°	2.1	53
N6-N5	24:22	.00137°	2.1	48
N6-N3	8:6	.00417°	6.3	880
N6-N5	25:23	.00137°	2.1	43
N6-N5	26:24	.00137°	2.1	39
N6-N3	9:7	.00412°	6.2	670
N6-N5	27:25	.00137°	2.0	35
N6-N5	28:26	.00137°	2.0	32
N6-N3	10:8	.00409°	6.1	530
N6-N5	29:27	.00137°	2.0	29
N6-N5	30:28	.00136°	2.0	26
N6-N3	11:9	.00406°	6.1	420
N6-N5	31:29	.00136°	2.0	24
N6-N5	32:30	.00136°	2.0	22
N6-N3	12:10	.00404°	6.1	340
N6-N4	6:4	.00579°	8.7	1700
N6-N5	33:31	.00136°	2.0	20
N6-N3	13:11	.00403°	6.0	270
N6-N5	34:32	.00136°	2.0	19
N6-N3	14:12	.00401°	6.0	230
N6-N5	35:33	.00136°	2.0	17
N6-N3	15:13	.00400°	6.0	190
N6-N5	36:34	.00136°	2.0	16
N6-N3	16:14	.00399°	6.0	160
net capture probability			~ 76%	

jectories? Tidal evolution brings the satellites into a resonant configuration, yet it can also carry them right through a resonance unless the strength of the resonance is enough to maintain the configuration. At very high rates of tidal dissipation, or with very weak resonances, the satellites' trajectories will not be remain in the resonance (Dermott *et al.* 1988). We can address this issue by examining the strongest tidal dissipation that will allow the resonance to remain intact at the point when the resonance is weakest. The strength of the resonance is proportional to the square of the inclination of 1989N6, so that at very small inclinations the resonance is very weak. Thus, if we find that the resonance is strong enough at plausible initial inclinations, the resonant configuration will be maintained throughout the satellite's tidal evolution and climb in inclination, ignoring other effects. A complete discussion of the mathematics involved can be found in M&D90. We have chosen to represent the minimum strength of the resonance by the minimum tidal Q_N value that would allow it to retain the satellites:

$$Q_N \geq 12 k_2 \frac{(p+2)}{b_{3/2}^{p+1}(\alpha)} \left(\frac{R_N}{a_x}\right)^5 \frac{a_x}{a_6} \frac{1}{i_6^2} \frac{1}{[p^2(\frac{a_x}{a_6})^2 + (p+2)^2 \frac{m_6}{m_x}]} \tag{1.6}$$

where m is a satellite mass and $b_{3/2}^{p+1}(\alpha)$ is the Laplace coefficient. The Love number is $k_2 = 3GM_N J_2/\omega^2 R_N^3 \approx 0.39$ (Dermott *et al.* 1988) using the values of J_2, M_N and R_N given in Tyler *et al.* (1989) and ω_N from Warwick *et al.* (1989). The results of this calculation for each of the resonances are also shown in Table II. These minimum Q_N's are all well under 12,000, which is likely to be the actual minimum value of Q_N as discussed above. Therefore, all of these primary resonances are strong enough to be viable.

We now need to determine which of the 35 possible resonances could have raised 1989N6 to its currently observed 4.7° inclination. T&W89 and M&D90 found that Miranda was ejected from an inclination resonance by a secondary resonance inside the primary one. This occurred at the 4.3° of inclination at which we find Miranda. Therefore, we will investigate the inclinations 1989N6 would be at if it had been ejected from each of these 35 primary resonances by a low-order secondary resonance. A good discussion of the events that occur during evolution in the resonance appears in T&W89, particularly their Figs. 5 through 14. These authors have described the system in terms of action-angle coordinates, which we adopted as well. The angle is chosen to be the argument of the i_6^2 resonance, which librates about a particular value if in resonance. The action is then a function directly related to the satellite's inclination. So, under tidal evolution in the resonance, the resonance is said to climb in inclination (action) and carry the satellite with it.

How do the satellites escape the resonance? As we mentioned before, the resonances are weakest at low inclinations, so that as they climb in inclination, they become stronger. This means that they occupy a larger region of space in action-angle coordinates. The strength of the $i_6 i_x$ resonance is proportional to the inclination of 1989N6 as well. So as 1989N6 climbs in inclination, the $i_6 i_x$ resonance is growing in strength too. At some point, these resonances will have grown in strength enough so that they are no longer well separated. That is, the libration regions of the resonances will nearly overlap, and the resonances will significantly perturb one another. This is dramatically evident in Fig. 7 of T&W89. At this point, the separatrices of

the resonances become chaotic, and secondary resonances, caused by perturbations from the $i_6 i_x$ resonance on the librations in the i_6^2 resonance, start to appear. The secondary resonances are the result of low-order commensurabilities between the libration frequency of the satellites' trajectories in the i_6^2 resonance with the perturbation frequency of the other resonances; in our case, the most important resonance is the $i_6 i_x$ resonance. The frequency of libration in the i_6^2 resonance is a function of the amplitude of the libration, being highest for very small oscillations and dropping to zero at the separatrix. Furthermore, the libration frequency for small oscillations increases as the strength of the resonance grows and the satellite climbs in inclination. The perturbation frequency of the $i_6 i_x$ resonance is given by the difference of the circulation rates of the resonant arguments of the i_6^2 and $i_6 i_x$ resonances. The result of all of this is that secondary resonances are created at the stable fixed point of the primary resonance and then move out towards the separatrix of the resonance as evolution continues. Thus, the secondary resonances sweep past the trajectory of the satellites in action-angle coordinates and can possibly capture their trajectory, much like the primary resonance did. If the secondary resonance does capture the satellites' trajectory, it then drags it out to the separatrix, where it can escape the primary resonance entirely.

To decide which primary resonance carried 1989N6 to its current inclination, we now need to examine each primary resonance's secondary resonances in similar detail as we did the primary ones. That is, their capture probabilities, strengths against tidal forcing, and inclinations at which they would eject the satellites from the primary resonance must be computed. This is because the scenario requires the satellites' trajectory to be caught into a sec-

ondary resonance (hence the necessity of the capture probabilities) and follow it out to the separatrix (meaning that stability against tides is necessary). At that point, the satellites' trajectory can escape the primary resonance at a particular inclination depending upon which primary and secondary resonances are being examined. This inclination of escape should match the observed 4.7° for the appropriate primary-secondary resonance combination responsible for 1989N6's inclination.

To compute the secondary resonances' capture probabilities, we followed the analysis presented in Malhotra (1990) for perturbations on the i_6^2 resonance by the $i_6 i_x$ resonance. She considers only the 4:1, 3:1, 2:1 and 1:1 resonances because those are the strongest secondary resonances. We find that this is reasonable for our case as well, since the 4:1 secondary resonances are already of almost negligible capture probability. Furthermore, we did not need to consider the 1:1 secondary resonances at all because in all cases, the 2:1 resonances were encountered below their critical value and thus all had a capture probability of unity. As a result, the trajectories could never reach the 1:1 secondary resonances. In Table III, we present the results of the calculations for capture probability into the 4:1, 3:1 and 2:1 secondary resonances for each of the possible primary resonances. For the capture probabilities in the table, we include both the probability concerning capture into that particular primary resonance, and the probability of not being caught into a secondary resonance before the one considered. That is, the probabilities listed in the table are the overall probabilities for the satellites' trajectory to be caught into that particular secondary resonance of that primary resonance. In these calculations, the initial inclination of 1989N6 was again

Table III: 1989N6 Secondary Resonance Results

Resonance		4:1 Secondary			3:1 Secondary			2:1 Secondary		
Sats	$p+2:p$	P(%)	Q_{min}	i_{escape}	P(%)	Q_{min}	i_{escape}	P(%)	Q_{min}	i_{escape}
N6-N5	16:14	0.3	3.3×10^4	2.0°	0.3	9.0×10^3	2.5°	1.5	260	3.8°
N6-N5	17:15	0.3	2.5×10^4	1.6°	0.4	7.6×10^3	2.0°	1.3	240	3.0°
N6-N3	6:4	$i_{encounter} > 15°$			$i_{encounter} > 15°$			$i_{encounter} > 15°$		
N6-N5	18:16	0.4	1.9×10^4	1.3°	0.4	6.4×10^3	1.7°	1.3	220	2.5°
N6-N5	19:17	0.4	1.6×10^4	1.1°	0.4	5.5×10^3	1.4°	1.3	200	2.0°
N6-N2	4:2	$i_{encounter} > 15°$			$i_{encounter} > 15°$			$i_{encounter} > 15°$		
N6-N5	20:18	0.4	1.3×10^4	0.9°	0.4	4.7×10^3	1.2°	1.2	190	1.7°
N6-N5	21:19	0.4	1.1×10^4	0.8°	0.4	4.1×10^3	1.0°	1.2	180	1.5°
N6-N3	7:5	$i_{encounter} > 15°$			$i_{encounter} > 15°$			$i_{encounter} > 15°$		
N6-N4	5:3	$i_{encounter} > 15°$			$i_{encounter} > 15°$			$i_{encounter} > 15°$		
N6-N5	22:20	0.5	8.7×10^3	0.7°	0.5	3.6×10^3	0.9°	1.2	160	1.3°
N6-N5	23:21	0.5	7.3×10^3	0.6°	0.5	2.3×10^3	0.8°	1.1	120	1.1°
N6-N5	24:22	0.5	6.2×10^3	0.5°	0.5	2.8×10^3	0.7°	1.1	140	1.0°
N6-N3	8:6	0.5	7.9×10^5	11°	0.9	1.1×10^5	15°	$i_{encounter} > 15°$		
N6-N5	25:23	0.5	5.3×10^3	0.4°	0.5	2.5×10^3	0.6°	1.1	140	0.9°
N6-N5	26:24	0.5	4.6×10^3	0.4°	0.5	2.3×10^3	0.5°	1.0	130	0.8°
N6-N3	9:7	0.5	4.4×10^5	7.0°	1.0	7.0×10^4	9.3°	4.7	1200	15°
N6-N5	27:25	0.6	4.0×10^3	0.4°	0.5	2.1×10^3	0.5°	1.0	120	0.7°
N6-N5	28:26	0.6	3.5×10^3	0.3°	0.5	1.9×10^3	0.4°	1.0	120	0.6°
N6-N3	10:8	0.6	2.6×10^5	4.8°	1.0	4.9×10^4	6.2°	4.5	960	9.7°
N6-N5	29:27	0.6	3.1×10^3	0.3°	0.5	1.7×10^3	0.4°	0.9	110	0.6°
N6-N5	30:28	0.6	2.8×10^3	0.3°	0.5	1.6×10^3	0.3°	0.9	110	0.5°
N6-N3	11:9	0.7	1.6×10^5	3.4°	1.1	3.5×10^4	4.3°	4.3	800	6.6°
N6-N5	31:29	0.6	2.5×10^3	0.2°	0.5	1.5×10^3	0.3°	0.9	100	0.5°
N6-N5	32:30	0.7	2.3×10^3	0.2°	0.5	1.4×10^3	0.3°	0.9	96	0.4°
N6-N3	12:10	0.8	1.0×10^5	2.4°	1.1	2.6×10^4	3.1°	4.1	690	4.7°
N6-N4	6:4	$i_{encounter} > 15°$			$i_{encounter} > 15°$			$i_{encounter} > 15°$		
N6-N5	33:31	0.7	2.1×10^3	0.2°	0.5	1.3×10^3	0.3°	0.9	93	0.4°
N6-N3	13:11	0.9	7.0×10^4	1.8°	1.2	2.0×10^4	2.3°	3.9	560	3.4°
N6-N5	34:32	0.7	2.0×10^3	0.2°	0.5	1.2×10^3	0.2°	0.8	90	0.4°
N6-N3	14:12	1.0	4.9×10^4	1.4°	1.2	1.6×10^4	1.7°	3.8	520	2.6°
N6-N5	35:33	0.7	1.8×10^3	0.2°	0.5	1.1×10^3	0.2°	0.8	87	0.3°
N6-N3	15:13	1.1	3.6×10^4	1.1°	1.3	1.2×10^4	1.3°	3.6	460	2.0°
N6-N5	36:34	0.7	1.7×10^3	0.2°	0.5	1.0×10^3	0.2°	0.8	84	0.3°
N6-N3	16:14	1.2	2.7×10^4	0.8°	1.3	1.0×10^4	1.1°	3.5	410	1.6°

assumed to be 0.06°. Some of the entries are listed as "$i_{encounter} > 15°$." The inclinations of encounter for these particular secondary resonances were sufficiently high that one of the central approximations of the analysis, that $\sin i \sim i$, was no longer valid.

A method for estimating the strength of the secondary resonances against tidal forcing is presented in M&D90. We have used this to compute the minimum Q_N values at which each of the secondary resonances could still capture a satellite. These calculations are also presented in Table III. Note that the minimum Q_N's of the differently indexed secondary resonances each vary by roughly an order of magnitude. The 2:1 resonances only need Q_N greater than about 400 in order to hold the trajectory, while the 3:1 resonances need Q_N greater than about 10^3. The 4:1 secondary resonances require Q_N values greater than about 10^4 to remain intact against tidal forces. We found earlier that due to considerations of tidal evolution and the location of co-rotation, the Q value of Neptune could be no less than about 12,000. Therefore, the 2:1 and 3:1 secondary resonances are unaffected by the tides, but the 4:1 resonances could have been broken up by them if the Q value of Neptune is near its lower limit.

The only question left to address is the actual inclination at which these trajectories escape from the resonance. This is the observable that we are trying to match with our scenario, and thus an important quantity. Previous works of this nature, e.g., T&W89 and M&D90 have performed numerical simulations of the evolution in the resonance to determine at which inclination the trajectory would escape. We calculate the inclination at which the satellite is ejected from these various resonances analytically. When

the primary resonances' libration regions start to overlap, their perturbations on one another become significant. In addition to creating secondary resonances, the perturbations also cause the separatrices of the primary resonances to broaden into regions of chaotic trajectories. Soon after a secondary resonance, bearing the satellites' trajectory, reaches the inner edge of the chaotic separatrix, the trajectory escapes from the primary resonance entirely. Therefore, we examine when the relevant secondary resonance meets the chaotic separatrix of the primary resonance, and then postulate that it is at this inclination that the trajectory will escape. We verify this approach by comparing with the results of numerical simulations carried out by M&D90.

The width of the separatrix can be estimated using the Melnikov-Arnold integral as outlined in Chirikov (1979), and Lichtenberg and Lieberman (1983). The libration amplitude at the inside edge of the primary resonance's chaotic separatrix, ϕ_{sx} is expressed by:

$$\cos\phi_{sx} = \frac{\Delta H}{H_{sx}} = 16\pi\frac{i_x}{i_6}\left(\frac{S}{\omega_{l_o}}\right)^2 \exp\left(-\frac{\pi S}{2\omega_{l_o}}\right) \tag{1.7}$$

where S is the perturbation frequency of the $i_6 i_x$ resonance on the i_6^2 resonance, $\omega_{l_o}(i)$ is the central libration frequency of the primary resonance, and $\Delta H/H_{sx}$ is the fractional variation of the energy along the separatrix of the primary resonance. We find the position of the secondary resonance by solving the equation for the libration frequency in the primary resonance as a function of libration amplitude, and equating that to a multiple of the $i_6 i_x$ resonance perturbation frequency. The perturbation frequency is a constant, and the libration frequency as a function of the libration amplitude, ϕ_{max}, is

given by (e.g., Lichtenberg and Lieberman 1983, p.26):

$$\omega_l = \omega_{l_o}(i)\frac{\pi}{2\mathcal{K}(\kappa)} \tag{1.8}$$

where $\kappa \equiv \sin \phi_{max}/2$, and $\mathcal{K}(\kappa)$ is the complete elliptic integral of the first kind. The inclination at which the secondary resonance and the chaotic separatrix first touch is then used as the inclination of escape. The results of these calculations for each of the secondary resonances in each of the primary resonances, are presented in Table III.

We have assumed that when the secondary resonance encounters the stochastic layer about the primary's separatrix, the trajectory escapes the primary resonance. We checked this assumption using statistics from M&D90. They ran 55 trajectories through similar resonances involving Miranda, and found distributions for the inclinations of escape. These are plotted in Fig. 1.3, along with a calculation of the width of the chaotic layer and the position of the secondary resonances in libration amplitude. The intersection between the inner edge of the chaotic zone and the location of the secondary resonance corresponds quite well with the inclination at which the trajectories escaped, verifying our assumption. Therefore, we feel that this calculation is a valid method for estimating the inclination at which a secondary resonance will eject a trajectory from a primary resonance. However, it is important to note that the M&D90 integrations were performed at an artificially low value of Q. This may have affected their results in that the stronger tides might have destroyed the actions of other subtle effects that were neither observed in their investigation nor accounted for in our

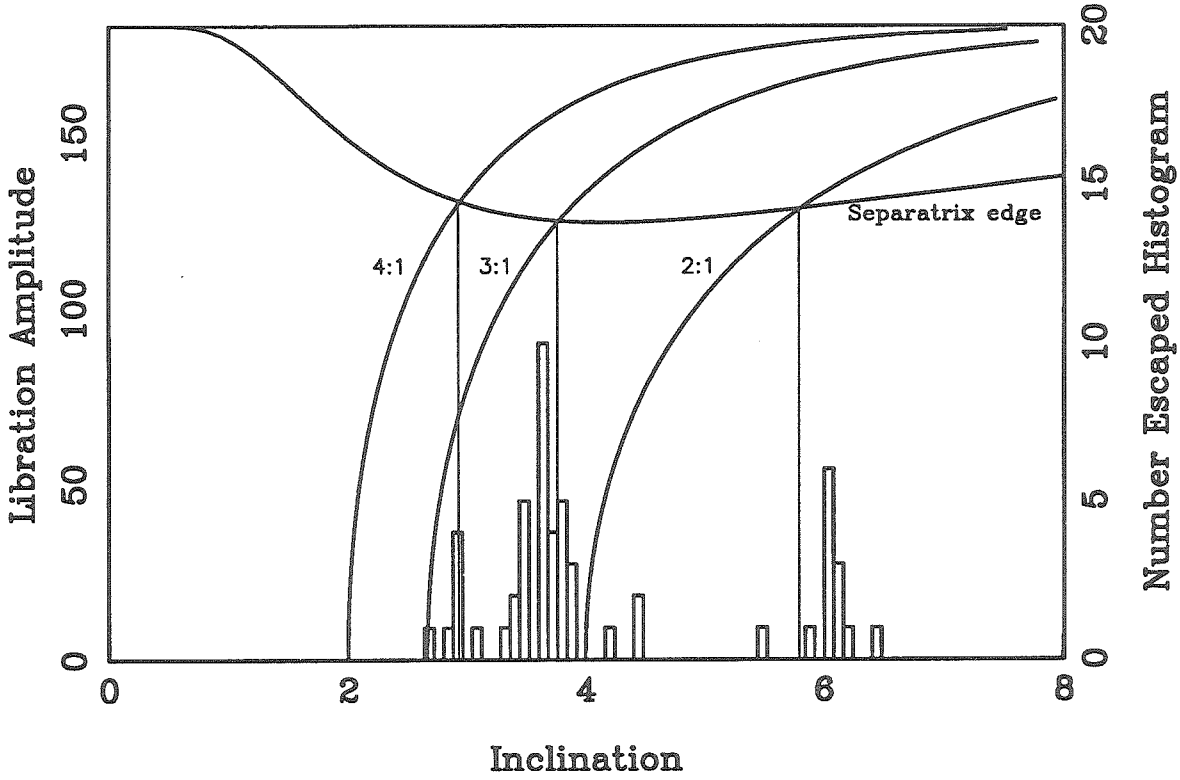


Fig. 1.3: This depicts the libration amplitudes that the 2:1, 3:1, and 4:1 secondary resonances can be found at as a function of inclination for the Miranda problem addressed by T&W89 and M&D90. It also shows the advance of the inner edge of the chaotic separatrix of the primary resonance. We have used the intersection of these curves to predict the inclination at which satellites captured in resonance would escape that resonance. To back this up, we show the results of numerical simulations by M&D90 in which they evolve 55 trajectories through the primary resonance. The inclinations of escape that they found agree quite well with the intersections of the curves

analyses.

Upon examining Table III, it is clear that there are 3 candidate resonances that could have resulted in an expulsion of 1989N6 from a primary inclination resonance at $i \sim 4.7°$. These are the 1989N6–1989N3 12:10 resonance (through its 2:1 secondary resonance), the 1989N6–1989N3 11:9 resonance (through its 3:1 secondary resonance), and the 1989N6–1989N3 10:8 resonance (through its 4:1 secondary resonance). The last one, employing the 4:1 secondary resonance has a probability of occurrence of at most 0.6%, assuming the Q value of Neptune is greater than 250,000; an assumption that is not wholly sound. The second one, using the 3:1 secondary resonance has a probability of occurrence of about 1.1%, assuming that Neptune's Q is greater than about 34,000. The first one mentioned, the 12:10 resonance between 1989N6 and 1989N3 with the orbit trajectory exiting the resonance on the back of the 2:1 secondary resonance has a probability of occurrence of about 4.1%. Therefore, we believe that this is the best candidate for explaining the inclination of 1989N6. The total probability, including capture into the primary resonance, capture into the secondary resonance, and escape at the right inclination works out to be $\sim 4\%$. This is not an overwhelmingly likely probability, but neither is it astronomically small. In fact as mentioned above, the more interesting statistic to look at is the probability that 1989N6 didn't get caught into any of these resonances, only 24%. Therefore, since it is very probable that 1989N6 was caught into one of the resonances, this particular resonance, being as likely as any of the others, is a good candidate. Furthermore, this scenario matches the observations.

Because the actual masses of the satellites are so poorly known, we did

these calculations again for both higher and lower ratios of the other satellites' masses to 1989N6. The results showed very similar capture probabilities and resonance strength limits on Q_N. The major difference was that the index of the primary resonance responsible for the inclination may have altered by one. That is, instead of the 12:10 resonance being the most likely candidate, the 11:9 (or 13:11) resonance through the 2:1 secondary resonance may have better fit the 4.7° observation. The important thing to note is that the probability of capture is unchanged, and the scenario is fundamentally unchanged as well. Therefore, the actual masses of the satellites, within the bounds consistent with the observations, are not critically important pieces of information to these calculations.

1.6 Other Satellites' Resonances

As a consistency check, it is possible to perform all of the resonance calculations for 1989N5 as well, treating it as we did 1989N6 above. One finds that 1989N5 passed at most 23 commensurabilities. If Q_N is taken to be at its lower limit, 12,000, then the probability that 1989N5 was caught into a resonance that would eject it at an inclination greater than what it is observed to be at $(0.14 \pm 0.06°)$ is about 70%. However, if Q_N is taken to be near its upper limit, 330,000, then the reduced tidal evolution means that 1989N5 passed fewer resonances, and had a probability of arriving at an inclination higher than that observed of only about 40%. Furthermore, regardless of the value of Q_N, there was about a 20% chance that 1989N5 was caught into a resonance and ejected at an inclination consistent with the observations.

Therefore, the resonances involving 1989N5 are not inconsistent with the observations, however they do slightly favor a high value of Q_N.

We mentioned above that there are in fact three inclination resonances for each mean motion commensurability. In the preceding discussions, we only considered the first one encountered, the i_6^2 resonance. We should recognize that the other two resonances *are* encountered, and if the trajectory escapes the first resonance, the other two could possibly catch it, and raise the other satellites' inclinations. We have calculated the capture probabilities for these other two sets of inclination resonances for each commensurability. The overall probability that one of these resonances caught either 1989N5 and another satellite or 1989N6 and another satellite is about 60%. This probability is calculated assuming Q_N is at its lower limit. If Q_N were near its upper limit, this probability would drop to about 40%. Because we do not observe the other satellites to be inclined to any great degree, we conclude that none of these resonance passages resulted in capture. This part of the analysis also somewhat favors a high value for Q_N.

The orbits of 1989N1-1989N4 are all diverging, so that no two of these satellites were ever captured into resonance. However, the inclination resonances are still encountered, and a consequence of a non-capture resonance encounter is that the satellite's inclination is slightly increased. The inclination increase is roughly that value corresponding to the critical inclination for capture into the resonance (Dermott *et al.* 1988). These satellites passed through something less than 21 non-capture inclination resonances. We have calculated the rough magnitude of the inclination increases this would result in for these satellites, finding $\Delta i \sim 0.05°$. This number is consistent with the

current distribution of inclinations of the inner satellites, which vary from about 0.014° to 0.2°. Although the current inclinations of 1989N1-1989N4 are consistent with these inclination changes, they can give no diagnostic support to the dynamical history of the inner satellite system.

Just as there are three inclination resonances for each commensurability, there are also three eccentricity resonances. We would like to be able to determine whether the inner satellites were caught into some of these resonances also, but this is difficult. Eccentricities damp in these satellites on a timescale of order 10^8 years, and therefore any memory of their eccentricity is lost after that time. Furthermore, because of this loss of memory, it is impossible to know at what eccentricity the satellites encountered the resonances, and thus impossible to know their capture probabilities. Nevertheless, if we naïvely assume that the eccentricities have always been roughly what they are now, we find that it is almost certainly true that both 1989N5 and 1989N6 were once captured into eccentricity resonances. This is consistent with observations because, as mentioned above, the satellites would quickly lose their memory of these resonance passages.

1.7 Conclusion

We have extended the scenario put forth in GMLB concerning the capture of Triton and its effects on the Neptune system to include the inner satellites discovered by Voyager. We note that 1989N1-1989N6 are all confined within $5R_N$ as predicted in GMLB. By showing that a satellite system existing at the time of Triton's capture would have destroyed itself due to mutual

collisions, we infer that the observed inner satellites formed on equatorial orbits after the Triton capture event and later cometary disruption events. The four innermost satellites are inside the Roche limit. There are several other examples of this in the Solar System. Hence, it is possible that these satellites accreted within the Roche limit. It is also possible that the innermost satellites, 1989N5 and 1989N6, are the largest shards of a parent body that migrated into the Roche zone under tides before its destruction, while 1989N2-1989N4 may be the tidally evolved remnants of other parent bodies which had also started outside the Roche limit.

The issue of 1989N6's 4.7° inclination is then resolved by examining the possibility of a temporary inclination resonance capture. We show that this is very likely (76%), in addition to being the only feasible mechanism. We suggest that 1989N6 was inclined to its observed value by the 1989N6-1989N3 12:10 resonance, ejected via the primary resonance's 2:1 secondary resonance at roughly $i \sim 4.7°$. This particular scenario has an overall probability of about 4%, comparable to or greater than all the other possible resonances. Finally, by applying tidal constraints, Q limits for Neptune are established. The lower limit is found by requiring that 1989N1 not pass corotation, and the upper limit by requiring the satellites, 1989N3 and 1989N6, to have once been in the 12:10 resonance. These limits for Q_N are:

$$12,000 < Q_N \times \frac{T_{circ}}{4 \times 10^9 years} < 330,000.$$

1.8 References

BORDERIES, N., AND P. GOLDREICH 1984. A Simple Derivation of Capture Probabilities for the $J+1 : J$ and $J+2 : J$ Orbit-Orbit Resonance Problems. *Celestial Mechanics* **32**, 127-136.

BURNS, J. A., M. R. SHOWALTER, AND G. E. MORFILL 1984. The Ethereal Rings of Jupiter and Saturn. In *Planetary Rings* (R. Greenberg and A. Brahic, Eds.), pp. 200-272.

CHIRIKOV, B. V. 1979. A Universal Instability of Many-Dimensional Oscillator Systems. *Physics Reports* **52**, 263-379.

COLWELL, J. E. , AND L. W. ESPOSITO 1992. Origin of the Rings of Uranus and Neptune: I. Statistics of Satellite Disruptions. *JGR Planets* submitted.

DERMOTT, S. F. , R. MALHOTRA, AND C. D. MURRAY 1988. Dynamics of the Uranian and Saturnian Satellite Systems: A Chaotic Route to Melting Miranda? *Icarus* **76**, 295-334.

FARINELLA, P., A. MILANI, A. M. NOBILI, AND G. B. VALSECCHI 1980. Some Remarks on the Capture of Triton and the Origin of Pluto. *Icarus* **44**, 810-812.

GOLDREICH, P., N. MURRAY, P. Y. LONGARETTI, AND D. BANFIELD 1989. Neptune's Story. *Science* **245**, 500-504.

GOLDREICH, P., AND S. SOTER 1966. Q in the Solar System. *Icarus* **5**, 375-389.

HARRIS, A. W. 1984. The Origin and Evolution of Planetary Rings. In *Planetary Rings* (R. Greenberg and A. Brahic, Eds.), pp. 641- 659. Univ. of Arizona Press, Tucson.

LICHTENBERG, A. J., AND M. A. LIEBERMAN 1983. *Regular and Stochastic Motion.* Springer-Verlag, New York.

P. Y. LONGARETTI 1989. Saturn's Main Ring Particle Size Distribution: An Analytical Approach. *Icarus* **81**, 51-73.

MALHOTRA, R., AND S. F. DERMOTT 1990. The Role of Secondary Resonances in the Orbital History of Miranda. *Icarus* **85**, 444-480.

MALHOTRA, R. 1990. Capture Probabilities for Secondary Resonances. *Icarus* **87**, 249-264.

MCKINNON, W. B. 1984. On the Origin of Triton and Pluto. *Nature* **311**, 355-358.

OWEN, JR., W. M., R. M. VAUGHAN, AND S. P. SYNNOTT 1991. Orbits of the Six New Satellites of Neptune. *AJ* **101**, 1511-1515.

S. J. PEALE 1986. Orbital Resonances, Unusual Configurations and Exotic Rotation States among Planetary Satellites. In *Satellites* (J. A. Burns and M. S. Matthews, Eds.), pp. 159-223. Univ. of Arizona Press, Tucson.

SMITH, B. A., L. A. SODERBLOM, D. BANFIELD, C. BARNET, A. T. BASILEVSKY, R. F. BEEBE, K. BOLLINGER, J. M. BOYCE, A. BRAHIC, G. A. BRIGGS, R. H. BROWN, C. CHYBA, S. A. COLLINS, T. COLVIN, A. F. COOK II, D. CRISP, S. K. CROFT, D. CRUIKSHANK, J. N. CUZZI, G. E. DANIELSON, M. E. DAVIES, E. DE JONG, L. DONES, D. GODFREY, J. GOGUEN, I. GRENIER, V. R. HAEMMERLE, H. HAMMEL, C. J. HANSEN, C. P. HELFENSTEIN, C. HOWELL, G. E. HUNT, A. P. INGERSOLL, T. V. JOHNSON, J. KARGEL, R. KIRK, D. I. KUEHN, S. LIMAYE, H. MASURSKY, A. MCEWEN, D. MORRISON, T. OWEN, W. OWEN, J. B. POLLACK, C. C. PORCO, K. RAGES, P. ROGERS, D. RUDY, C. SAGAN, J. SCHWARTZ, E. M. SHOEMAKER, M. SHOWALTER, B. SICARDY, D. SIMONELLI, J. SPENCER, L. A. SROMOVSKY, C. STOKER, R. G. STROM, V. E. SUOMI, S. P. SYNOTT, R. J. TERRILE, P. THOMAS, W. R. THOMPSON, A. VERBISCER, AND J. VEVERKA 1989. Voyager 2 at Neptune: Imaging Science Results. *Science* **246**, 1422-1449.

STEVENSON, D. J., A. W. HARRIS, AND J. I. LUNINE 1986. Origins Of Satellites. In *Satellites* (J. A. Burns and M. S. Matthews, Eds.), pp. 39-88. Univ. of Arizona Press, Tucson.

THOMAS, P., AND J. VEVERKA 1991. Neptune's Small Inner Satellites. *JGR Planets* **96**, 19261-19268.

TITTEMORE, W. C., AND J. WISDOM 1988. Tidal Evolution of the Uranian Satellites. I. Passage of Ariel and Umbriel through the 5:3 Mean

Motion Commensurability. *Icarus* 74, 174-230.

TITTEMORE, W. C., AND J. WISDOM 1989. Tidal Evolution of the Uranian Satellites. II. An Explanation of the Anomalously High Orbital Inclination of Miranda. *Icarus* 78, 63-89.

TYLER, G. R., D. N. SWEETNAM, J. D. ANDERSON, S. E. BORUTZKI, J. K. CAMPBELL, V. R. ESHELMAN, D. L. GRESH, E. M. GURROLA, D. P. HINSON, N. KAWASHIMA, E. R. KURSINSKI, G. S. LEVY, G. F. LINDAL, J. R. LYONS, E. A. MAROUF, P. A. ROSEN, R. A. SIMPSON, AND G. E. WOOD 1989. Voyager Radio Science Observations of Neptune and Triton. *Science* 246, 1466-1473.

WARWICK, J. W., D. R. EVANS, G. R. PELTZER, R. G. PELTZER, J. H. ROMIG, C. B. SAWYER, A. C. RIDDLE, A. E. SCHWEITZER, M. D. DESCH, M. L. KAISER, W. M. FARRELL, T. D. CARR, I. DE PATER, D. H. STAELIN, S. GULKIS, R. L. POYNTER, A. BOISCHOT, F. GENOVA, Y. LEBLANC, A. LECACHEUX, B. M. PEDERSEN, P. ZARKA 1989. Voyager Planetary Radio Astronomy at Neptune. *Science* 246, 1498-1501.

Chapter 2

A Steady-State Kalman Filter for Assimilating Data from a Single Polar Orbiting Satellite

2.1 Abstract

We examine a steady-state scheme for data assimilation in the context of a single, short period (relative to a day), sun-synchronous, polar-orbiting satellite. If the satellite takes observations continuously, the gains, which are the weights for blending observations and predictions together, are steady in time. For a linear system forced by random noise, the optimal steady-state gains (Wiener gains) are equivalent to those of a Kalman filter. Computing the Kalman gains increases the computational cost of the model by a large factor, but computing the Wiener gains does not. The latter are computed

by iteration using prior estimates of the gains to assimilate simulated observations of one run of the model, termed 'Truth,' into another run termed 'prediction.' At each stage, the prediction errors form the basis for the next estimate of the gains. Steady-state is achieved after 3 or 4 iterations. Further simplification is achieved by making the gains depend on longitudinal distance from the observation point, not on absolute longitude. For a single-layer primitive equation model, the scheme works well even if only the mass field is observed but not the velocity field. Although the scheme was developed for Mars Observer, it should be applicable to data retrieved from Earth atmosphere satellites, e.g., UARS.

2.2 Introduction

Kalman filtering is a recipe for blending predictions of a model with observations of a physical system to obtain an optimal estimate of the current system state (e.g., Gelb 1974). Unfortunately, Kalman filtering increases the computational burden of the model by a factor of order M, the number of state variables in the model system. This burden is prohibitive for atmospheric general circulation models (GCMs), where M is 10^5 or greater. In addition Kalman assimilation is often numerically unstable for large systems (Bierman 1977), so Kalman filtering has been an unreachable goal for atmospheric GCMs (Ghil *et al.* 1981, Miller 1986, Ghil 1989, Cohn and Parrish 1991, Daley 1992b).

The problem arises because the prediction error covariance matrix, whose dimension is $M \times M$, must be advanced forward in time as the system evolves.

This time-dependence arises because the observing pattern, the observation error, and the model error can all be time-dependent. But when these quantities (and thus the prediction error covariance matrix) are constant in time, the Kalman filter approaches a steady-state. In such a case, the computational burden can be reduced to a factor of order one, i.e., the model runs almost as fast with data assimilation as without it because the prediction error covariance matrix need not be advanced in time. Moreover, the performance of such a steady-state Kalman filter is optimal, at least for linear constant-coefficient systems forced by white noise. Steady-state Kalman filters (Wiener filters) have much in common with Optimal Interpolation (OI; see Bengtsson and Gustavsson 1971, Rutherford 1972, Bergman 1979), which is much used in operational forecasting. But currently operational observing systems have little in common with a system comprised of only a single polar orbiting satellite. The present paper is an attempt to describe and evaluate an application to a single-satellite observing system.

Knowledge of the prediction error covariance matrix is necessary for computing the optimal weights (gains) that are used for blending the predictions and the observations together. Our iterative scheme uses the current best estimate of the gains to assimilate data during a simulated flight of the spacecraft, from which a new estimate of the prediction error covariance matrix is obtained. The simulation uses one run of a model as 'Truth' and another run for prediction. Simulated observations, with realistic errors, are made from the 'Truth' run. The resulting data are then assimilated into the prediction run using the current best estimate of the gains. The errors are measured by subtracting the 'Truth' field from the prediction field, and the error products

are collected as sums. After integration for a suitable length of time, the sums are turned into averages and a new estimate of the prediction error covariance matrix is obtained. This leads to a new estimate of the gains, and the iteration is repeated. Convergence is complete in three or four iterations. The steady-state gains are then used with real data for as long as the observing pattern and the statistical properties of the system remain constant in time.

There are advantages and disadvantages to using a model as 'Truth.' An advantage is that the model's 'Truth' is known, whereas that of the real atmosphere is not. Knowing the 'Truth' allows us to bypass costly forward-integration of the prediction error covariance matrix: Instead of trying to compute the errors, we can measure them, which requires much less computer time. The disadvantage is that the model 'Truth' is not the real truth, which is the atmosphere itself. Since we are using a model both for prediction and for 'Truth,' our estimates of the error would tend to be low. To counter this identical twin problem, we add system noise to the 'Truth' run but not to the prediction run. The result provides an approximation to the true model error.

Our scheme was developed for use with Mars Observer. Although the spacecraft failed before going into orbit, the goal was to produce a continuous record of global weather over the planned two-year lifetime of the mission. The Mars GCM (MGCM) developed at Ames Research Center (Pollack *et al.* 1990, Haberle *et al.* 1992) was chosen as the model. Like all GCMs, it computes the state vector (temperature, pressure, wind, dust, and water vapor on a finite-difference global grid) at time $t + \Delta t$ from that at time t.

Our plan was to assimilate data continuously, i.e., to use observations taken during each integration step of the model to modify the state vector at that time step. The amount of modification takes into account the expected error of the observation and the expected error of the prediction. The output of the model after the assimilation step becomes the current best estimate of the global weather. The model then integrates this estimate forward to the next time step, during which new observations are collected and the process is repeated.

Mars Observer was designed for a sun-synchronous polar orbit with a period of 0.08 sols (Martian days). While Mars spins, the orbit stays fixed at the same local time of day, so successive dayside equator crossings move west by slightly less than 30 degrees. In one sol the spacecraft covers the globe with 12-13 equally-spaced (in longitude) polar passes on both dayside and nightside. For our purposes, a steady-state observing system is one where the same instruments are taking data from the same point in the orbit on each polar pass. Our first objective was to assimilate data from one instrument, the Pressure Modulator Infrared Radiometer (PMIRR), which obtains profiles of temperature, water vapor, and dust to 80 km altitude every 115 km down-track, i.e., along the track of the spacecraft (McCleese *et al.* 1992). Since the spacecraft moves at 3.3 km/s, more than 10 profiles are obtained during each 7-minute time step of the MGCM.

The foundations of our assimilation technique are expanded upon in the next section, showing its roots in standard OI theory and its relation to Kalman and Wiener filters. We discuss the assumptions that we have made in the details of our assimilation technique. We also introduce a method of

determining our gain functions, which we call Wiener gains. We have experimented with a number of simple (and not so simple) models to verify our assumptions. The remainder of the paper is devoted to detailing those models and the conclusions drawn from them. We first use a linear Rossby wave model that has one dimension (longitude) and one dynamical variable (horizontal streamfunction). We demonstrate the effectiveness of our technique of deriving the Wiener gains on this model, and discuss the east-west asymmetry of the gain function for a sun-synchronous satellite. We then use a 2-D one-variable model to show that the gain functions can be made functions of latitude as well as latitudinal and longitudinal distance between the spacecraft and the analysis point. With this model, we also investigate the severity of ignoring the serial correlations of observation errors. The next model is a three-variable Rossby wave/gravity wave model in one dimension. Only one variable, geopotential, is observed. The wind vector is analyzed from its correlation with geopotential. Furthermore, with this model we test whether topography is a serious perturbation to the Wiener gains, and thus whether we can ignore it. Finally, we implement our scheme on a non-linear spherical shallow water model, tuned to simulate Mars. The full implementation on the MGCM will be described in a later paper.

2.3 Assimilation Gains

2.3.1 Background Theory

The key to data assimilation lies in determining the weighting of the observations in computing an analysis. These weights, or gain functions, specify the changes to be applied to the prediction at an analysis point as a function of the difference between the prediction and observation at the observation point. This is mathematically written as

$$\Psi_k^a = \Psi_k^p + \sum_{j=1}^{N} \alpha_{kj}(\Psi_j^o - \Psi_j^p) \tag{2.1}$$

where Ψ_k is an element of the state vector (at a point k) and the α_{kj}'s are the gain function coefficients specifying how the discrepancy between the observation and prediction at point j should affect the state variables at point k. There are N observations in this update. The superscripts a, p and o represent the analysis, prediction and observation state vectors respectively.

Optimal interpolation (OI) is based on minimizing the analysis errors that result from blending model predictions with observational data (Bengtsson and Gustavsson 1971, Rutherford 1972, Bergman 1979). OI tries to determine the values of α_{kj} that give the best estimate of the true state of the physical system. For a proper derivation of OI, the reader is referred to Rutherford (1972). The assumptions that one makes in the definition of OI are that the physical system and the observing system are statistically steady and that the observational errors are uncorrelated with the prediction errors. Minimizing the errors made in the analysis state vector yields a matrix equa-

tion for the gain functions as a function of the observation error covariance and the prediction error covariance,

$$\sum_{j=1}^{N}(\overline{\epsilon_i^p \epsilon_j^p} + \overline{\epsilon_i^o \epsilon_j^o})\alpha_{kj} = \overline{\epsilon_k^p \epsilon_i^p} \tag{2.2}$$

where the ϵ_i^p's are prediction errors at point i and the ϵ_i^o's are observation errors at point i. The overbars indicate averaging over many realizations. This equation is solved for the gain functions by inverting the matrix on the left side.

The above equations are expressed in a form that implicitly handles a large number of observations at once, assimilating them in parallel. For Earth, observations from the global synoptic network are typically assimilated at 0,6,12, and 18 hours GMT (e.g., Baker *et al.* 1987, DiMego 1988). Satellite observations are usually simply moved to the synoptic times where they are assimilated along with data from the synoptic network. For Mars, however, and to some extent for the Earth's upper atmosphere, the data are all asynoptic. Different longitudes are observed at different times. Our approach is to assimilate data serially, at the time step nearest to when they were taken.

If the observation errors are uncorrelated, i.e., the observation error covariance matrix $\overline{\epsilon_i^o \epsilon_j^o}$ is diagonal, then each observation can be handled separately. Equation (2.2), expressing the gain function for each observation,

then becomes a simple division:

$$\alpha_{ki} = \frac{\overline{\epsilon_k^p \epsilon_i^p}}{\overline{\epsilon_i^p \epsilon_i^p} + \sigma_i^2} \tag{2.3}$$

where σ_i^2 is the error variance for observation i. The assumption of no correlations in the observation errors is generally not valid (Daley 1992a), however, we will show later that making this assumption presents little loss in performance for an application like Mars Observer. Jazwinski (1970, p.198) discusses processing the components of the observation vector one by one when the observation error covariance is diagonal and there is no change in the state vector due to the dynamics. He states, "That this is equivalent to processing the whole observation vector at once is conceptually clear and can be proved algebraically (very tedious!)." Jazwinski does not give a proof, but we outline one in the appendix.

Equation (2.1) relates an observation at point j to the kth element of the state vector, Ψ_k. Efficient but sub-optimal forms of OI are obtained by assuming that the prediction error correlations are compact in space, symmetric about the observation point, and have Gaussian or other simple dependence on distance (e.g., Baker *et al.* 1987). In this paper k includes *all* elements, so that the assimilation step is global in extent. We make no assumptions about the spatial form of the prediction error covariance matrix. Further, we make no assumptions (e.g., geostrophic balance) about how different variables (e.g., wind and pressure) are related. Instead, we rely on the prediction error covariances to establish the relation.

The equivalence of these gains to the gains of a steady-state Kalman

filter is easily seen by examining the equation for the discrete Kalman gain function (expressed in matrix form):

$$\mathbf{K} = \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1} \tag{2.4}$$

where $\mathbf{K}$ is the Kalman gain function matrix, $\mathbf{P}\mathbf{H}^T$ is the prediction error covariance matrix times the transpose of the observation matrix and $\mathbf{R}$ is the observational error covariance matrix (Gelb 1974). This equation is simply a more general form of (2.2), where the matrix $\mathbf{H}$ contains the transform between the observations and the state vector elements. Thus, our technique differs from Kalman only in the way in which the prediction error covariance matrix is determined. Kalman filtering advances it mathematically in time with the state vector, while in our technique, it is measured via models and stored for later use with real data.

2.3.2 Gain Simplifications

In order that the gains be constant in time, the atmosphere and the observation system must be statistically steady. This means that we must have different gains for different seasons, and perhaps also for periods when the atmosphere is clear and periods when it is dusty. In addition, the characteristics of the instrument must not change, i.e., the observation error must be constant in time. And finally, the observing pattern must be steady.

The last condition needs some explanation. We need separate gains for each latitude of observation, because the coverage of the poles is different from that of the equator. For example, points near the poles are observed

every orbit, whereas points near the equator are observed once per day. But if the instrument is taking data all the time, then each time the spacecraft passes a certain latitude the positions of past observations relative to the spacecraft is the same. In other words, the observing pattern is constant. This satisfies the criterion for a steady-state Kalman filter. Even if the instrument is off during part of the orbit, the gains will be steady provided the on-off cycle is a function of latitude only.

In principle, the gains should also be functions of the longitude of observation. Relative to a scale height, the topography of Mars is significantly greater than the Earth's (Esposito *et al.* 1992), so the weather at one longitude is systematically different from that at another. Thus the prediction error covariance matrix should depend not only on the longitudinal distance from the observation point j to the analysis point k, but also on the absolute longitude of both. We have chosen to ignore this latter dependence. Below, we show that even for Mars, with its great topographic relief, the assumption presents an insignificant loss of performance. Thus, in ignoring longitudinal variations of the gain functions, we have reduced the dimensionality of the gain functions from four (2-D for all observing points and 2-D for all analysis points) to three (1-D for all observing latitudes and still 2-D for all analysis points). The magnitude of the problem of maintaining special gain functions is more manageable in this case.

We do not assume that the gain functions are symmetric about the observation point. For a polar-orbiting spacecraft, points to the east of the observation point were observed on the previous orbit, while points to the west were observed on the previous day. Thus, one would expect to find some

east-west asymmetry in the prediction error covariances of such an observing system. We observe such asymmetries in the results of some of our models, justifying this general treatment of the gain functions, although the asymmetry appears less important in other of our models. The experiments we present later in the paper exhibit the increase in performance in anisotropic over isotropic gain functions, and show that the loss in ignoring topography is minor. As a reference, we compare our assimilation scheme to Kalman filtering with favorable results. However, we also need to demonstrate a method of determining these Wiener gain functions.

2.3.3 Gain Function Recipe

As shown in (2.2) and (2.3), the gains α_{kj} depend on the error covariances. The observation error depends on instrument noise and on retrieval uncertainty — that associated with converting calibrated radiances into vertical profiles of physical quantities. The retrieval process introduces correlations within each profile, so the observational error should be a matrix whose dimension is the number of points in the profile. Such considerations are beyond the scope of this paper, however, since we are not discussing the implementation into the 3-D MGCM. In our simpler models presented herein, we assume that the observational errors are uncorrelated from one horizontal location to the next. The observational error variance is an input parameter to the assimilation scheme.

The other quantity that appears in (2.2) and (2.3) is the prediction error covariance, which we obtain by iteration. The first estimate comes from

the weather covariances, the spatial correlations of the weather fluctuations, which are the differences between the instantaneous state vector and its time mean. These covariances are much larger than the prediction error covariances obtained with the optimal gains, but they have a similar spatial structure.

The iteration process has been described earlier. With an estimate of the error covariances, one determines the gains using (2.2) or (2.3). One then assimilates some synthetic observations of the 'Truth' run into a prediction run, keeping track of the prediction errors. From these a new prediction error covariance is computed, and the process repeats. We do not mathematically show the convergence of this method, but have tested it on a number of different models, and all have converged in a few iterations. Furthermore, we have compared it to Kalman filters implemented on linear models, and the gain functions converge to the same functions as the Kalman gains in steady-state.

2.4 1-D One-Variable Model

In this section, we introduce the iterative procedure yielding the Wiener gains with an application to a 1-D Rossby wave model. We used a 1-D Rossby wave model on a latitude circle written in a geopotential-like variable, with damping and random forcing. The equation of motion is:

$$(\frac{\partial}{\partial t} + \frac{1}{\tau})(\frac{\partial^2}{\partial x^2} - \frac{1}{L_D^2})\Psi + \beta\frac{\partial \Psi}{\partial x} = F \tag{2.5}$$

where F is the random forcing, taken to be white noise in space and time, and β, L_D and τ were chosen to represent possible Mars-like conditions at the equator. x is longitude with period 2π, and t is time with one sol being 2π. L_D is the radius of deformation. We choose $L_D = 1/3$, about one third of Mars' radius in the model's dimensionless units. It emphasizes wavenumber three Rossby waves, like those reported in Barnes(1980, 1981) from the Viking Lander data. τ was taken to be 8π, corresponding to a decay time of 4 sols (Barnes 1980, 1981, Banfield *et al.* 1994). β is 2 in these dimensionless units. The equation yields a spectrum of Rossby waves. Wavenumbers 1,2 and 3 have the largest amplitudes and a coherence time of order 4 sols. We decomposed the system into 12 spectral modes of Fourier sine and cosine functions plus a constant term, with 25 terms in the representation. Observations of the state were degraded with Gaussian noise which had an amplitude of 5% of the average amplitude of the system. We modeled our observing system after Mars Observer, which would have crossed the equator on a descending node about 12 times a sol. Thus, the observing pattern was simply an observation about every 1/12 of a sol. This was also the model time-step as we analytically advanced the models in time from one observation to the next.

The simplicity of the system allowed us to *analytically* integrate the equation of motion from time t to time $t + \Delta t$, at which point the forcing changes discontinuously and a new observation is made. There is no leapfrog time step; the system remains first-order in time, and is computationally stable. We are thus able to derive explicit formulations for the transition and noise forcing matrices of the Kalman formalism. We implemented a Kalman filter

for this system, which is depicted as it begins to assimilate data in Fig. 2.1. The vertical bars in the 'Truth' model column are the actual observations made under the spacecraft, which is drifting westwards, making one cycle in a sol. They simply represent the value of the observed geopotential under the spacecraft, as measured from the dotted line which is the x-axis, with some artificial observational noise. The Kalman filter very quickly adjusts the estimate to the 'Truth' run in roughly a sol. Similarly, the shape of the Kalman gains can be seen to change over that time, as the prediction error covariances adjust to the new influx of information. The gain functions quickly settle to a steady-state which is a function of the longitudinal difference between the spacecraft and the analysis point only. In other words, the gain functions have a constant shape in the reference frame of the spacecraft.

Initially the Kalman gains are symmetric about the observation point and fairly broad. However, over the course of the next sol, they exhibit a shorter correlation length and more asymmetry from east to west. The asymmetry arises because the location just to the east of the spacecraft was visited one orbit (0.08 sols) earlier, while the location to the west of the spacecraft was visited 12 orbits (one sol) earlier. Additionally, β causes information to be advected westwards, adding to the asymmetry. The steady-state Kalman gain follows the spacecraft as it travels westward. It is depicted again as the solid line in Fig. 2.2. The asymmetry is particularly evident in this figure, where the spacecraft is taken to be at $x = 0$.

The analyses are all done in the spectral coordinates of the model, and thus the gain functions relate an observation in space to a correction of a spectral mode. The gains in Fig. 2.2 are obtained from the spectral coefficient

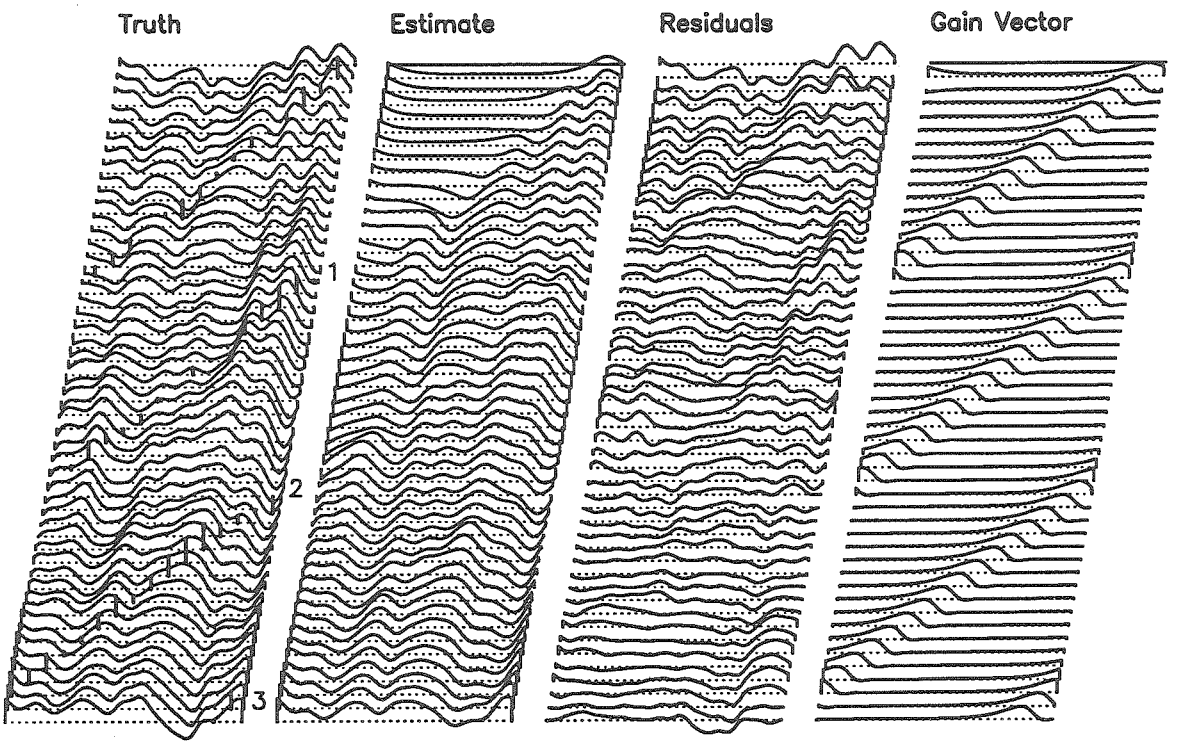


Fig. 2.1: 1-D 1-Variable model's evolution over about 3 sols, with time advancing downwards and each column representing one quantity over the full domain of the model. The leftmost column shows the state of the 'Truth' run, or what is taken to be the real atmospheric state. Observations, represented by the vertical bars from the dotted zero line to the curve, are being taken by a satellite which progresses westward with time. The second column corresponds to an estimate of the true state made by a Kalman filter. The third column shows the difference between the 'Truth' and the estimate. Finally, the fourth column shows the Kalman gains used to weight the observation taken at that time step. Note the speed with which the Kalman filter improves the estimate, and how the shape of the gain function changes over the first sol.

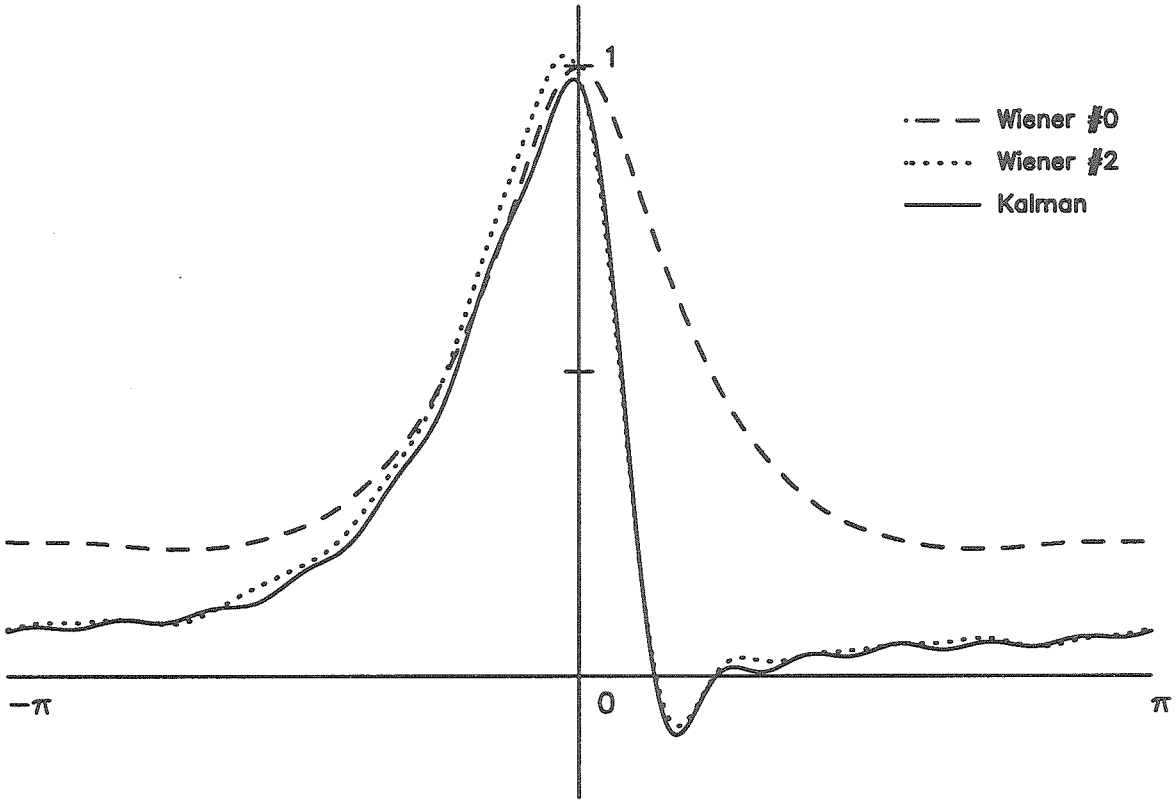


Fig. 2.2: Three gain functions for the 1-D 1-Variable model, the zeroth iterate Wiener gain, the second iterate Wiener gain and the steady-state Kalman gain. The spacecraft is at $x = 0$. A gain of 1.0 means that the observation is solely used to make the analysis, while a gain of 0.0 means that only the prediction is used in the analysis. The similarity between the second iteration Wiener gains and the Kalman gains shows the near optimal nature of the iterated Wiener gains.

gains by a linear transformation. This method is used throughout the paper, including the more complex models that follow. It is a natural choice for these models as they are formulated in spectral space, and it is currently in use in operational schemes (Parrish and Derber 1992).

The other two curves in Fig. 2.2 are the zeroth and second iteration of the Wiener gains for this system. The zeroth iteration is computed from (2.3) using the total weather covariance as the prediction error covariance. This is necessarily a sub-optimal gain function, as it is the equivalent of the prediction errors made using no predictive model at all. It does however roughly correspond to the western half of the Kalman gain function, perhaps indicating that that side of the Kalman gain is controlled by the correlation length scale of this model's weather. Since the zeroth iterate has no information about the spacecraft's orbit, it is a symmetric (i.e., isotropic) function. After some 1000 sols of model time, the statistics were clean enough on the prediction error covariances in this iteration to yield the new (first iterate) gain functions. This procedure was then repeated to generate the second iteration gain functions, the dotted line in Fig. 2.2. The latter are strikingly similar to the Kalman gains, and perform just as well as they do. Further iterations change the gain functions insignificantly from either the second iterate, or the Kalman gain. Therefore, we find that this iterative technique does in fact converge to the steady-state Kalman gains, or the true Wiener gains, for this simple model.

Figure 2.3 shows the predictive performance during the first 10 sols of assimilation using the Kalman gains, the zeroth iterate Wiener gains, and the second iterate Wiener gains. Random weather variations make these curves

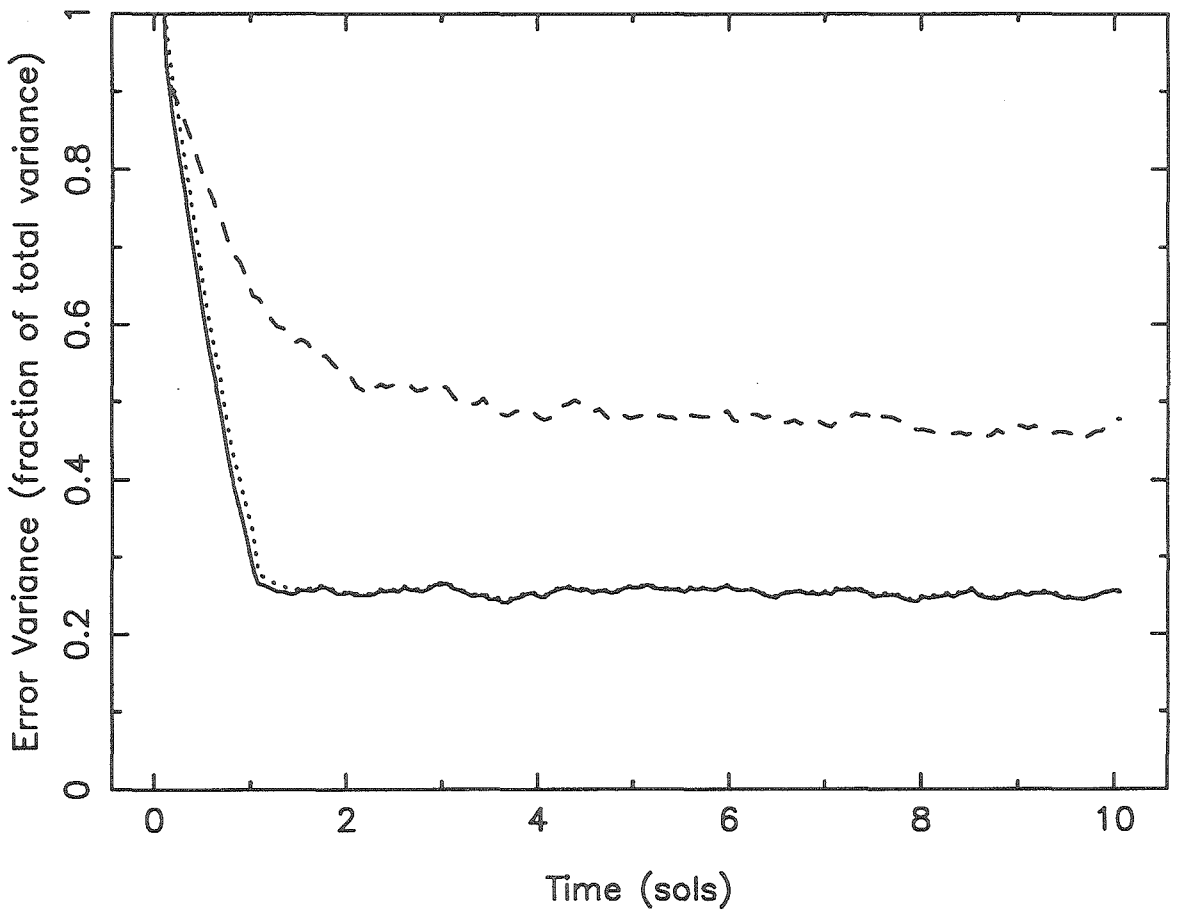


Fig. 2.3: Performance of the three gain functions of Fig. 2.2. The ordinate is the variance of the analysis errors from the model's true state. The error variances are normalized by the model's weather variance, so that an error variance of 1.0 is correctly analyzing essentially none of the model's weather variance. The curves run for 10 sols, and were all started from a zero state at the initial time. The dashed curve is for the zeroth iterate Wiener gains, which performs relatively poorly. The other two curves correspond to the iterated Wiener gains and the Kalman gains, which are nearly identical in performance and clearly superior to the zeroth iterate Wiener gains.

much less smooth for any particular 10 sols, but this plot is the result of averaging 1000 independent 10 sol start-up periods for the different filters. The prediction models were initialized to a zero state for all of these runs, so the initial errors were of the same order of magnitude as the model weather variance. The errors are normalized to the field's average variance, a normalized error of 1.0 corresponding to essentially no accuracy in the analysis, and a normalized error of 0.0 meaning that the predictive model exactly tracks the 'Truth' model's trajectory. Again, the solid curve corresponds to using the Kalman gains, which were able to bring the error variance down to about 25% of the model's weather variance in just over one sol. The dashed line corresponds to using the zeroth iterate Wiener gains, which took two or three sols to reach a steady-state with an error variance of about 46%. The dotted line corresponds to using the second iterate Wiener gains, the performance of which is almost indistinguishable from the Kalman gains.

This is a good model for comparing the computational burdens of the Kalman filter versus the constant Wiener filter. Every sol of assimilation using the Kalman filter on this model took three times as long as the Wiener filter to execute. The Kalman filter has the extra overhead of propagating the prediction error covariance matrix through every time step. For more complex models, this step becomes prohibitively slow because it scales as M^2, the square of the number of elements in the state vector. In our models, the Wiener filter's computational requirements scale as M, the same as in OI. Therefore, we expect our technique to present about the same computational burden as OI, and significantly less than Kalman filtering.

2.5 2-D One-Variable Model

To test the validity of this technique in a model with varying weather statistics from north to south, we used a 2-D one-variable Rossby wave model. This system is described by :

$$(\frac{\partial}{\partial t} + \frac{1}{\tau})(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} - \frac{1}{L_D^2})\Psi + \beta\frac{\partial \Psi}{\partial x} = F \tag{2.6}$$

where the only differences from the 1-D model are that the state vector now consists of the components of a 2-D Fourier transform. We used somewhat lower resolution for this model to reduce the computational burden; zonal wavenumbers of zero to seven, and meridional wavenumbers of one to five satisfying a no-normal-flow boundary condition. We implemented a Kalman filter for this model and also derived the iterated Wiener filters. The spacecraft orbit was modeled as traveling once per sol around the planet to the west, and making about 12 north-south passes in that time. The gain functions were kept in the reference frame of the spacecraft as separate functions for each latitude that the spacecraft observed, because the weather and prediction error covariances are assumed to depend on latitude only. Also, the gain functions were kept as *global* functions, relating the observation point to analysis points everywhere in the domain. This is the same as in the 1-D model, but with a 2-D domain and separate gain functions for different spacecraft latitudes. This implementation of the iterated Wiener filter worked just as well as the 1-D implementation, with similar improvements in performance between the zeroth iterate gain function and the later iter-

ates. Furthermore, the performance difference between the Kalman gains and the later iterations on the Wiener gains were indistinguishable. Finally, the shape of the Wiener gains was almost indistinguishable from the Kalman gains after only two iterations.

This model is also a good one with which to examine the impact of serially correlated observation errors. In this series of experiments, the observations were degraded with noise of the form $n_i = \rho n_{i-1} + (1 - \rho^2)^{1/2} r_i$, where n_i is the noise of the ith observation. ρ is a number less than 1, and r_i is a random number. The correlation coefficient between the noise at position i and that at position $i - m$ is ρ^m. The n_i have the same mean and variance as the r_i. We process these observations as before, ignoring the fact that the errors are correlated, and see what effect an increase in ρ has on the performance of the assimilation scheme. The serial correlations were modelled to have roughly the same meridional structure as the weather fluctuations themselves. For 10 observations in each north-south pass, this means a serial correlation of about 0.75 from one observation to the next. We modelled the northernmost observation as uncorrelated with observations preceding it.

Presented in Figure 2.4 is the performance of the zeroth, first, second and third iteration gain functions with an observation error of 5% relative to the weather fluctuations. The solid lines represent the performance without serially correlated observation errors $\rho = 0$, and the dotted lines represent those with serially correlated observation errors $\rho = 0.75$. The performance loss in ignoring these serial correlations is apparently negligible for this application, which in structure resembles the scenario of Mars Observer. The observation errors comprise such a small fraction of the prediction errors for

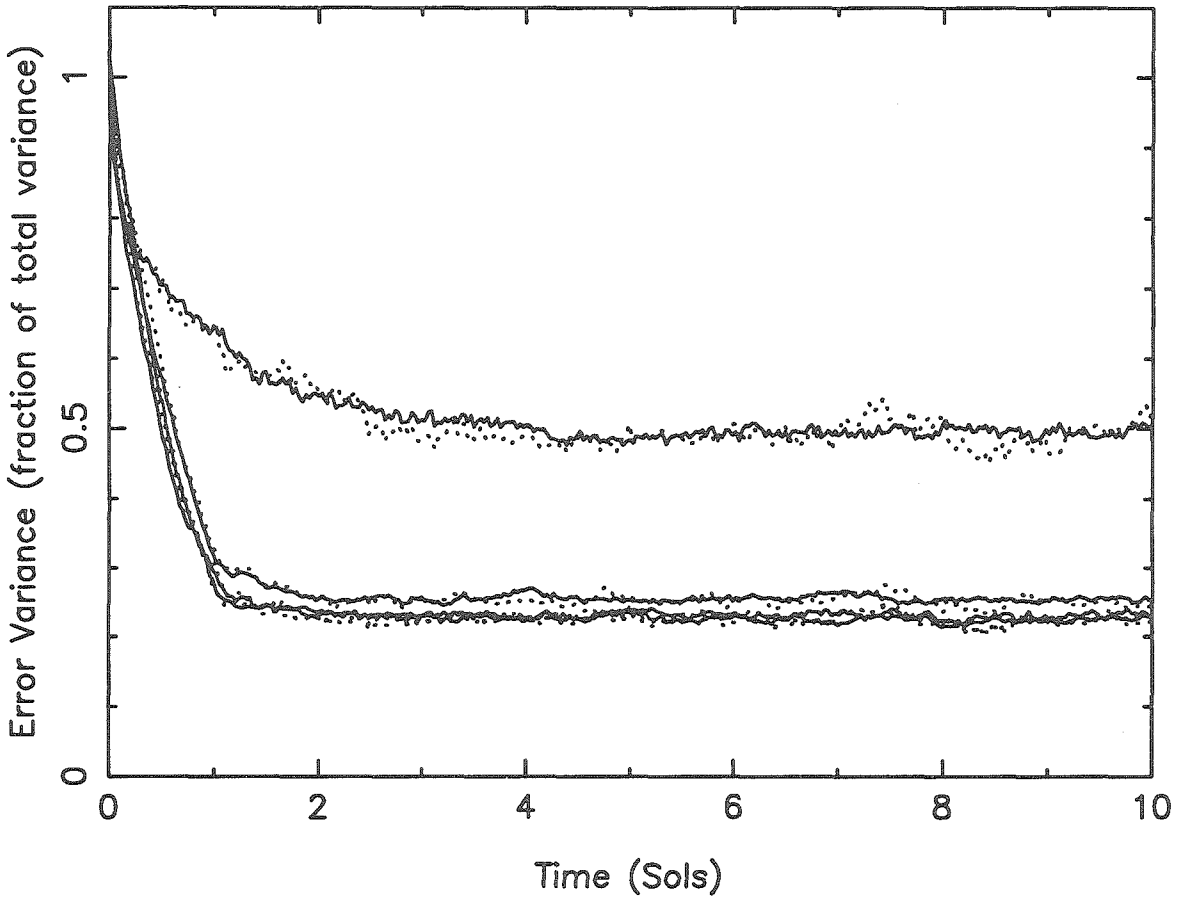


Fig. 2.4: Performance of the zeroth, first second and third gain functions for the 2-D 1-Variable model both with and without serially correlated observation errors. This is for observation errors of 5% of the total weather variance of the model. The solid lines represent the runs without serially correlated observation errors, and the dotted lines represent the corresponding runs with serially correlated observation errors. The error variance is normalized in the same way as in Fig. 2.3. The zeroth iterate yielded the solid and dotted lines asymptoting to about 0.5, while the subsequent iterates yielded the lower error variances. Note the similarity between each solid line and its corresponding dotted line.

this problem, that whether they are serially correlated or not is insignificant. However, if the observation errors were of greater amplitude, this is less true.

The PMIRR instrument of Mars Observer would have yielded atmospheric radiances which could have been inverted for temperatures with an accuracy of about 1-2K (McCleese *et al.* 1992). The atmospheric waves on Mars represent variations on the order of 5K (e.g., Barnes 1980). Therefore, a more reasonable estimate of the observation errors for the Mars Observer scenario is about 40%. Figure 2.5 shows the decrease in performance when serially correlated observation errors are present for the case of 40% errors. It shows a more notable decrease in performance than the 5% case, indicating that observation errors contribute more strongly to the prediction errors in this case. Nevertheless, the performance loss is small compared to the performance difference between the zeroth iterate gain function and subsequent iterates of the gains functions. Because of this, we believe that it is an adequate assumption to ignore the possibility of serial correlations in the observation errors for the Mars Observer scenario. However, if the observation errors are still greater, the assumption clearly will break down. Figure 2.6 shows the performance difference for a case with 200% observation errors. In this case, ignoring the serial correlations causes performance losses which are much greater than the performance increases found between the zeroth and subsequent iterations.

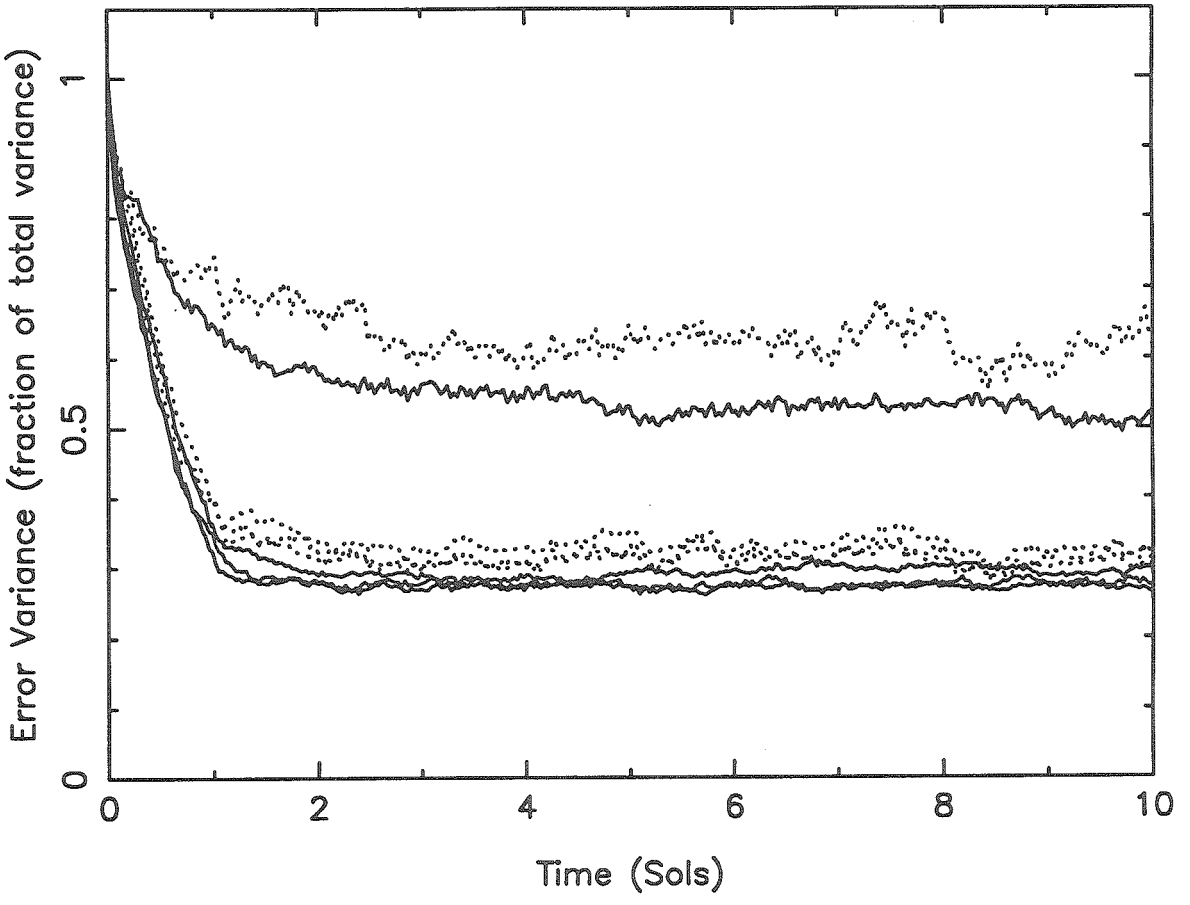


Fig. 2.5: Same as Fig. 2.4, but for observation errors of 40% of the total weather variance. This value is similar to that which had been expected for Mars Observer. Note that the later iterates still perform better than the zeroth iterate, even when there are serially correlated observation errors. This suggests that ignoring serial correlations of the observation errors does not represent a serious loss in performance.

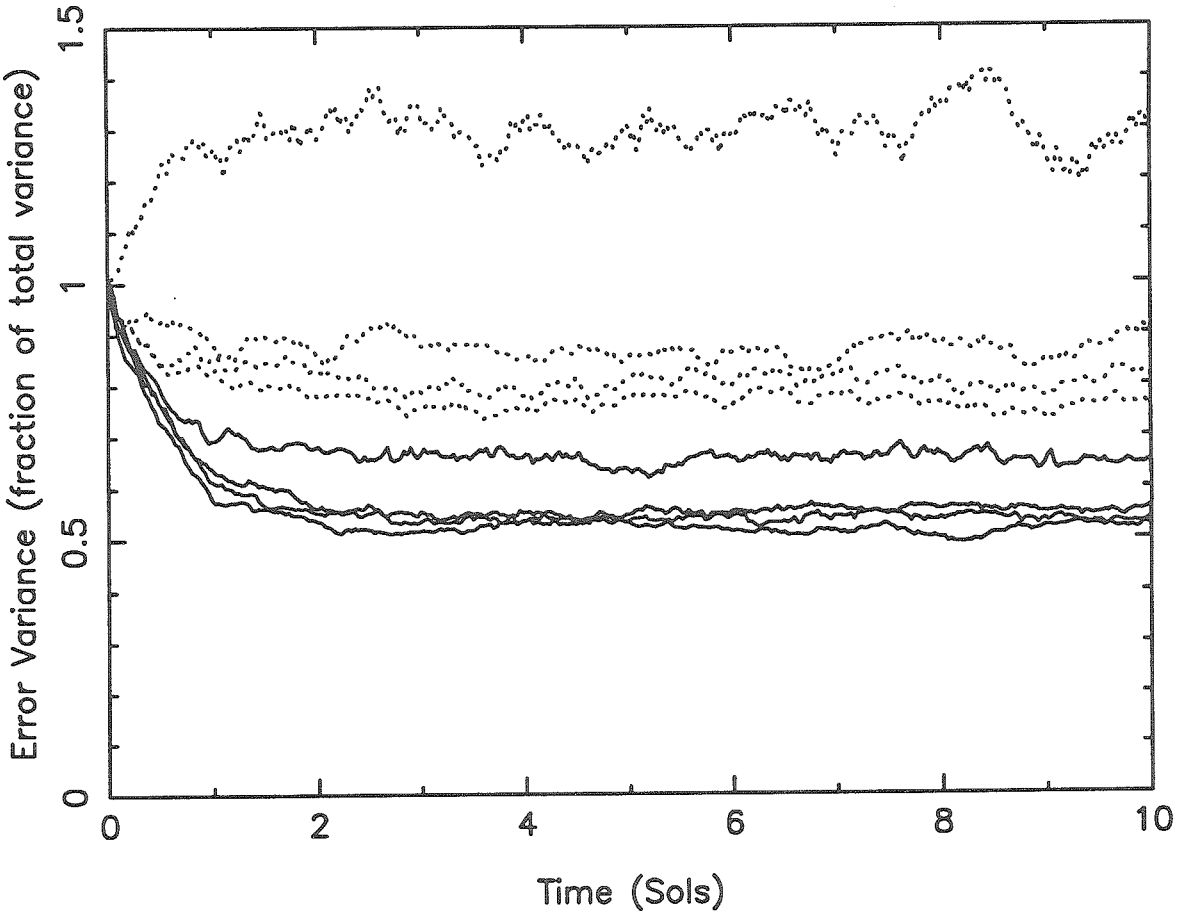


Fig. 2.6: Same as Fig. 2.4, but for observation errors of 200% of the total weather variance. In this extreme case, the runs with serially correlated observation errors perform significantly worse than those without. The zeroth iterate's performance is actually worse than no assimilation at all. In such a case, serial correlations of observation errors can not be ignored.

2.6 1-D Three-Variable Model

To test the performance of the Wiener gains on the primitive equation system, we applied them to the following linearized 1-D shallow water model. The objective was to observe only the mass field and see how well the mass and velocity fields follow the 'Truth' run.

$$\begin{aligned} (\frac{\partial}{\partial t}+\frac{1}{\tau})\Psi + f\chi &= F \\ (\frac{\partial}{\partial t}+\frac{1}{\tau})\chi - f\Psi + \Phi &= G \\ (\frac{\partial}{\partial t}+\frac{1}{\tau})\Phi + c^2\frac{\partial^2\chi}{\partial x^2} &= H \end{aligned} \tag{2.7}$$

In (2.7) Ψ is the streamfunction, χ is the velocity potential and Φ is the geopotential. τ is again a damping time ($\sim$ 4 sols), f is the coriolis parameter (taken to be at 30°N), and beta is neglected in order to simplify the algebra for Kalman filter implementation. c^2 is the critical wave speed, with a value of 1/9. With these settings, the radius of deformation, $L_D = c/f$, is again 1/3. The forcing functions, F, G and H are red in their time and space spectra (Lorentzians centered at zero frequency and zero wavenumber, with half-power points at a period of 4 sols and a longitudinal wavenumber of 3), so as not to force the gravity waves too strongly. The relative amplitudes of the forcings are chosen to mimic the non-linear terms which have been dropped from the system. Thus, $F \sim H$ and $G \sim F/4$, because the amplitude of χ is one- quarter that of Ψ and Φ when the forcing period is 4 sols. This model is

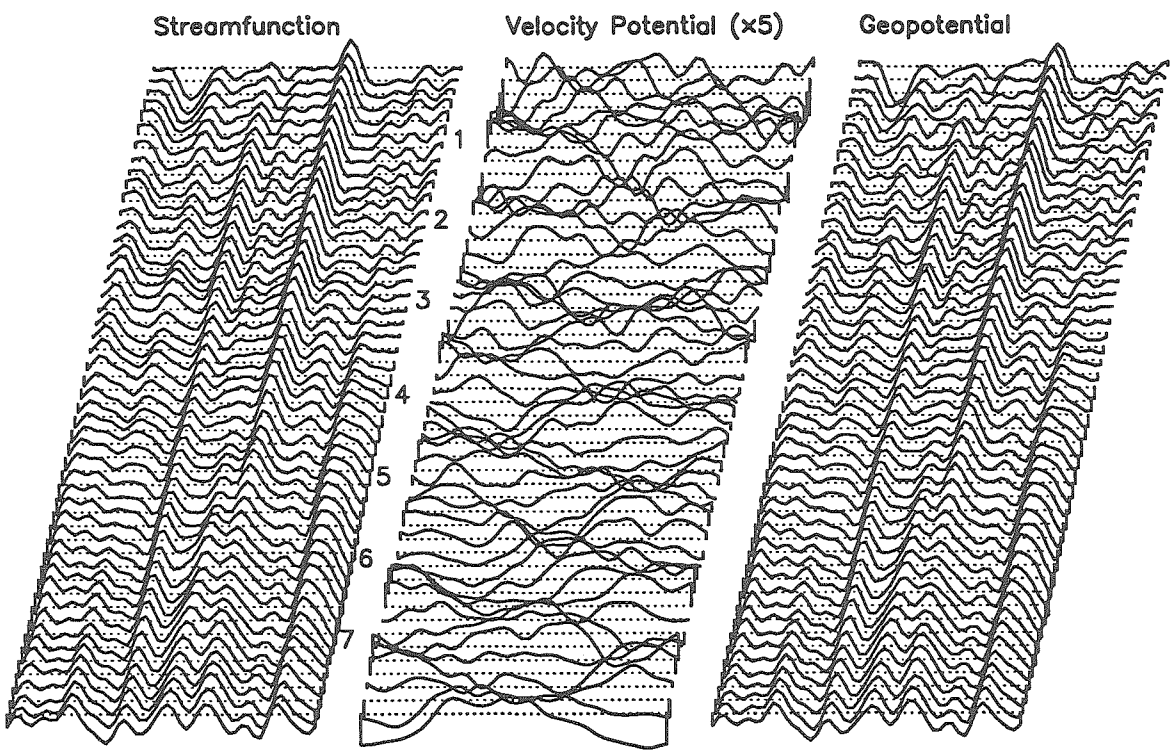


Fig. 2.7: Evolution of the 1-D 3-variable model over an 8-sol interval. Again, time advances downwards, and each column shows a dynamical quantity over the whole domain of the model. The streamfunction and geopotential evolve slowly and exhibit a great degree of geostrophy, while the velocity potential is dominated by quickly propagating gravity waves.

like the simple 1-D one-variable model with the addition that it also admits fast modes. The time evolution of the model's three variables during an eight sol run of this model is shown in Fig. 2.7. The geopotential and streamfunction fields are nearly in geostrophic balance. They also closely resemble the geopotential field of the simple one-variable model discussed above. However, the velocity potential field is dominated by rapidly propagating east

and west-bound gravity waves. It is smaller in amplitude than the other two fields by a factor of five, reflecting the fact that gravity waves are not strongly forced in this model and that geostrophic balance dominates. The dynamical result of neglecting β is that the slow modes do not propagate, although they drift randomly under the influence of the forcing. However, Rossby waves propagate rather slowly even when β is included, so we feel that this is not a crucial difference. Geopotential only is observed, and a 5% error has been added to the observations as before.

Again, we tested the iterative Wiener filter technique on this model and compared the results to a real Kalman filter. The results were similar to those obtained with the two Rossby wave models. After a few iterations, the Wiener gains were almost indistinguishable from the Kalman gains, and performed equally well. Figure 2.8 shows the Kalman gains, and the zeroth and fifth iteration Wiener gains. In this case there are three gain functions, because they relate an observation in geopotential to changes in all three dynamical variables. One can still notice the asymmetry between the east and west sides of the sub-spacecraft point first noticed with the 1-D one variable model. In this case, the asymmetry is not as dramatic, probably because this model has no β which likely serves to enhance east-west asymmetries.

The performance increase from the symmetric zeroth iterate gain function to the fifth iterate is not as drastic as with the simpler models, but still worth noting. Figures 2.9a, 2.9b and 2.9c depict the magnitude of the prediction error variance for the first 10 sols of assimilation for the streamfunction, velocity potential and geopotential fields respectively. The predictive runs were initialized to a zero state in all cases. The zeroth iterate (dashed line)

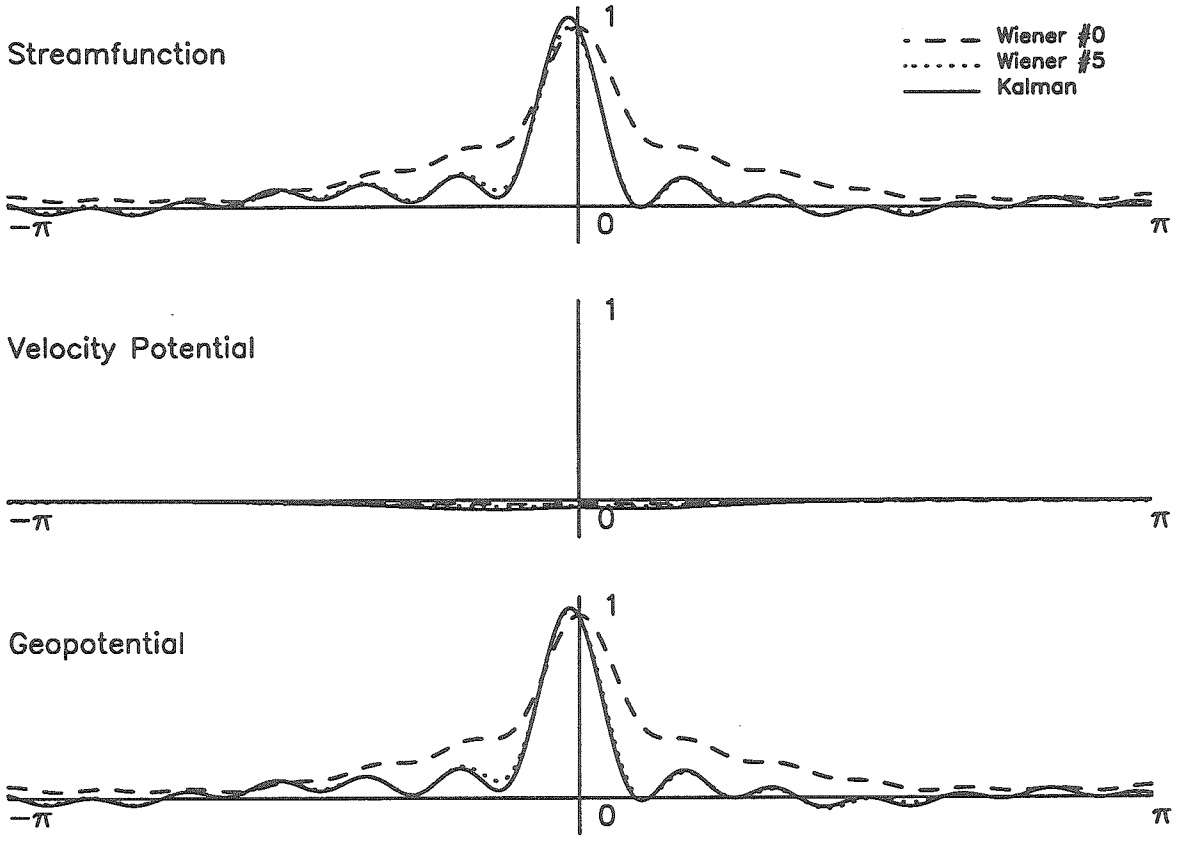
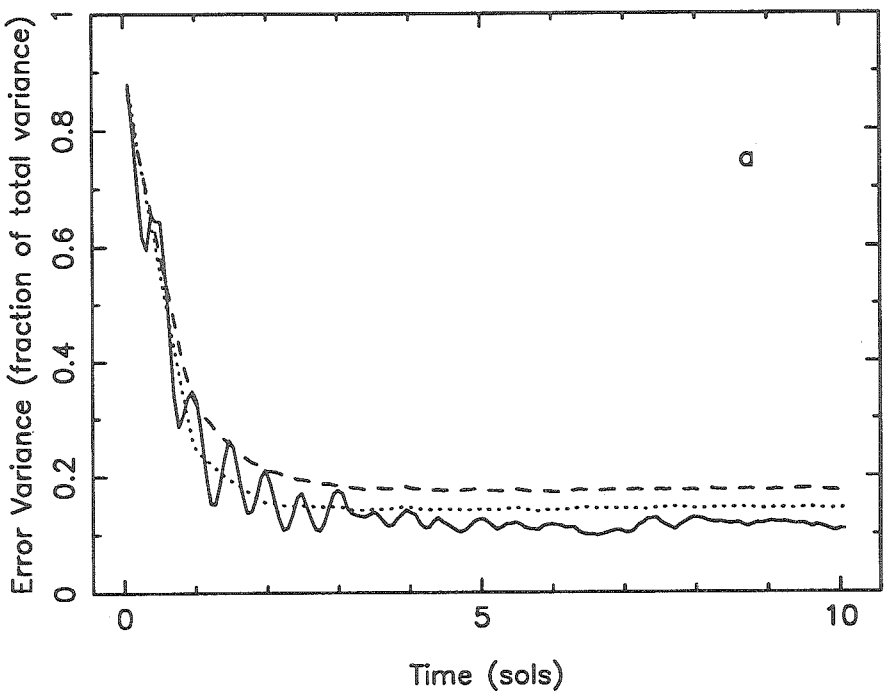
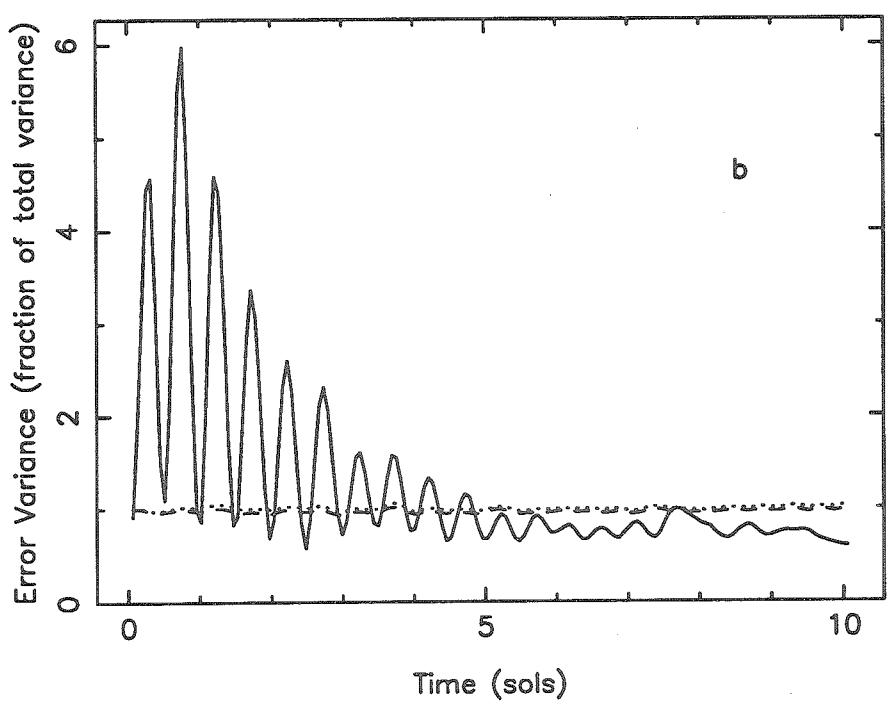


Fig. 2.8: Gain functions for the 1-D 3-variable model, each relating an observation of geopotential to an analysis weight in one of the dynamical variables. As in the 1-variable model, the iterated Wiener gains are nearly indistinguishable from the Kalman gains and longitudinally asymmetric. The gain functions for streamfunction and geopotential are similar due to the high degree of geostrophy present in the model. The low values of the velocity potential gain functions are indicative of the independent nature of geopotential and velocity potential in this model.


Error Variance (fraction of total variance)
0
0.2
0.4
0.6
0.8
1
a
0
5
10
Time (sols)


Error Variance (fraction of total variance)
0
2
4
6
b
0
5
10
Time (sols)

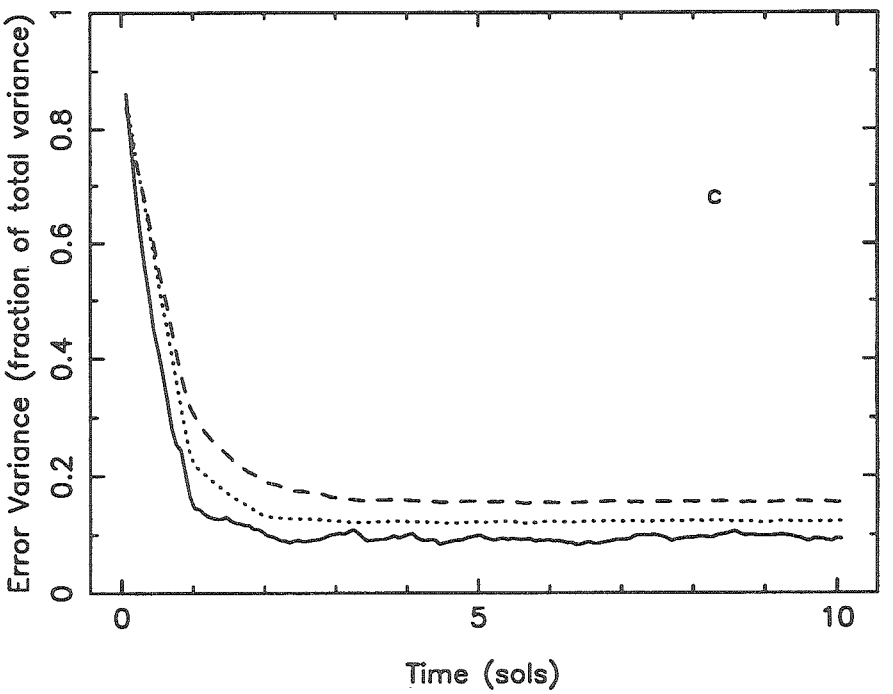


Fig. 2.9: Performance of the 3 sets of gain functions presented in fig. 2.8 for a period of 10 sols. The predictive runs were started from a zero state at the initial time. The three graphs show the analysis errors in streamfunction, velocity potential and geopotential, respectively. Note the speed with which the Kalman gains (solid line) reduce the errors, and the favorable comparison of the fifth iterate Wiener gains (dotted line) with the Kalman gains. Furthermore, note the increase in performance between the zeroth iterate and the fifth iterate Wiener gains in steady-state. Note the different vertical scale on 2.9b.

reaches a steady-state in about 3 sols, predicting all but 18% of the variance of the model in streamfunction, 16% in geopotential, and essentially nothing of the velocity potential field (> 98% unexplained variance). The fifth iterate Wiener gains (dotted line) reach a steady-state in only about 2 sols, predicting all but about 15% of streamfunction, 12% of geopotential, and again essentially none of velocity potential (> 98% unexplained variance). The Kalman filter (solid line) produces some strong oscillations at first, especially in the velocity potential field, probably as a result of improperly specified initial prediction error covariances. However, at least in the geopotential field, it settles down in about 2 sols. When the filter reaches its steady-state, it predicts all but about 9% of geopotential variance, about 11% of streamfunction variance and appears to predict a small percent of the velocity potential variance ($\sim$ 70% unexplained variance).

The fact that none of the sets of gain functions is able to effectively predict the velocity potential is not surprising because the only observed field is geopotential, which is dominated by Rossby waves, while the velocity potential is dominated by gravity waves. The lack of communication between these fields is also reflected in the much smaller amplitudes for the velocity potential gain function. The fact that streamfunction and geopotential gain functions are similar in shape and predictive ability simply reflects a high degree of geostrophy in the model.

This model is interesting in that it demonstrates how one can effectively use observations of one dynamical variable to predict the others. Cross-correlations between an observed variable and other dynamical variables are commonly used in operational assimilation schemes. It is these cross-

correlations that allow one to use information about one observed dynamical variable to guide the trajectory of the entire dynamical system through phase space. Mars Observer would have returned temperature soundings of the Martian atmosphere, and would have no access to direct wind measurements. It is then important to be able to predict the full dynamical state of the system from measurements of only one variable. This scheme, like currently operational OI schemes (e.g., Parrish and Derber 1992) makes no assumptions about dynamical balance in the cross-correlations relating different variables. Rather, we use the prediction error covariances relating the system variables with each other. This should give the right relation between the dynamical variables of the model system, without the difficulties geostrophy presents near the equator.

We neglected β in this model in order to be able to easily solve the equations of motion analytically, and thus easily cast the system into a Kalman filter formulation. However, in anticipation of real applications of this technique with finite difference models (GCMs), we implemented this model in a finite difference formulation as well. We used a leapfrog time-stepping scheme to advance the dynamics, thereby changing the system from first-order to second-order in time. The new state vector (at time $t + \Delta t$) is computed from the current and past state vectors (times t and $t - \Delta t$). We applied the same corrections to the past state vector as to the current one. Applying the correction to only the current state vector excited the computational mode in the integration.

Transforming this model into a finite difference model also allowed us to add β and topography to the model. We have been ignoring topography in

assuming that the gain functions are not functions of the absolute longitude of the spacecraft. Adding topography to the model is a way of checking this assumption. With these changes, the equations of the system become:

$$\begin{aligned}
(\frac{\partial}{\partial t}+\frac{1}{\tau})\Psi + f\chi &= F \\
(\frac{\partial}{\partial t}+\frac{1}{\tau})\chi - f\Psi + \Phi &= G \\
(\frac{\partial}{\partial t}+\frac{1}{\tau})\Phi + \frac{\partial}{\partial x}(c^2(x)\frac{\partial \chi}{\partial x}) - \beta L^2\frac{\partial \Phi}{\partial x} &= H
\end{aligned} \tag{2.8}$$

where $c^2(x)$ represents the topography through $c^2(x) = gh(x)$, gravity times the local depth of the layer. These expressions are not directly derivable from the shallow water equations, but they have the same behavior at both low and high frequencies. They have the further property of energy conservation in the absence of forcing and damping. We used a Galerkin scheme with a finite number of Fourier modes as the set of basis functions. This scheme conserves energy when the leapfrog time step is sufficiently small. We chose the function graphed in Fig. 2.10 as the layer depth. It is roughly 13km minus the meridionally averaged topography of Mars, scaled to yield a high degree of modulation on $c^2(x)$. Since the topographic relief of Mars is quite dramatic (on the order of an atmospheric scale height) this should be a good test of the relevance of specifying the gain functions in the spacecraft's frame of reference.

We include the effect of topography on the dynamics both for the 'Truth' runs, which have random forcing, and for the prediction runs, which are

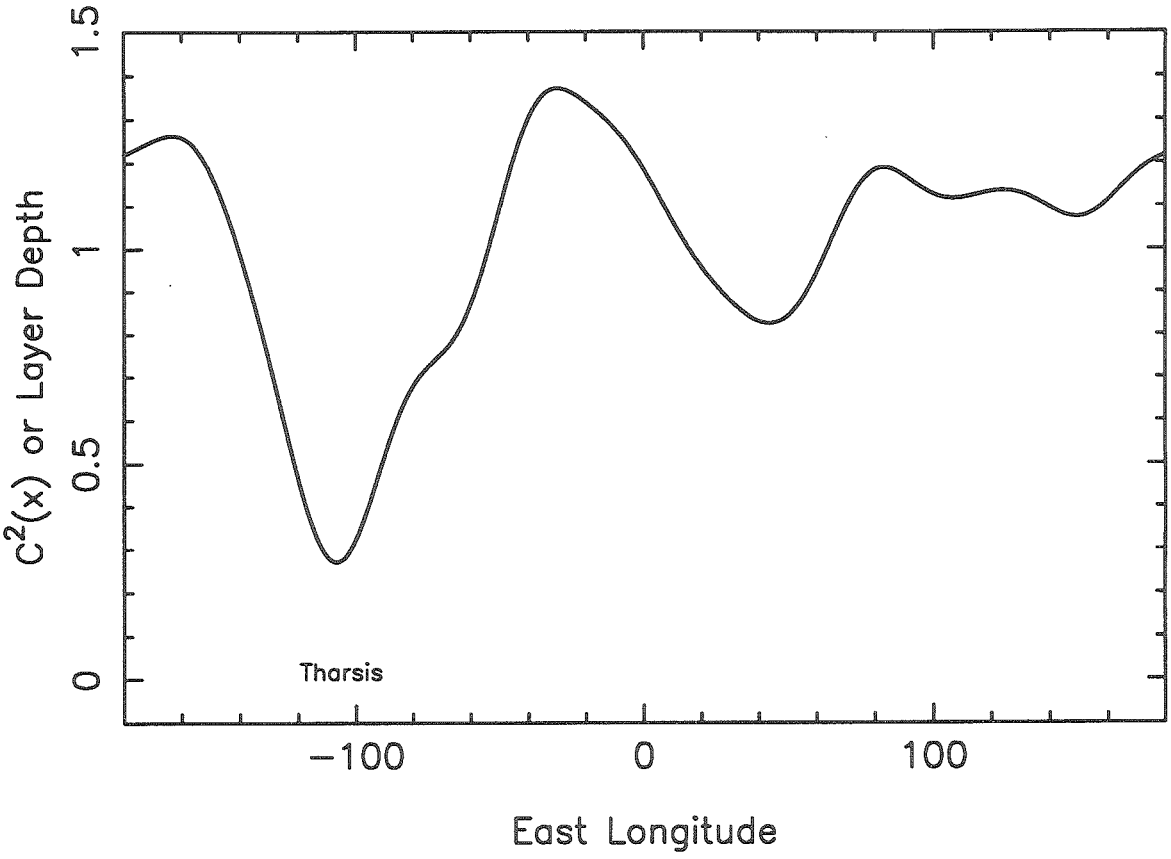


Fig. 2.10: Layer depth of the 1-D 3-variable model as a function of longitude. It corresponds to the actual Martian topography, subtracted from a reference of 13km, meridionally averaged and scaled to yield a high degree of modulation on the quantity c^2. The Tharsis plateau of Mars shows up quite clearly at around $-110°$. It appears as a low value because this is layer *depth*, as opposed to surface height.

forced only by assimilated observations. However, we ignore the effect of topography on the gain functions when we assume that they are the same for all spacecraft longitudes. With this assumption, we derived the longitudinally averaged iterated Wiener gains in the spacecraft's frame of reference and evaluated the predictive performance of these gain functions. Then, we included the effect of topography by having separate gain functions for the different spacecraft longitudes, and again derived a set of Wiener gains. These latter gain functions explicitly included information about the local topography, and thus should have performed better than the former gain functions.

The solid line in Fig. 2.11 represents the gain functions derived ignoring the presence of topography, assuming that the same function can be used for all spacecraft longitudes. The dashed line shows a gain function derived with topography explicitly included, one of a set which is different for each spacecraft longitude. This particular gain function is for the longitude with the greatest topographic relief, Mars' Tharsis region. As one might expect, this is the location where the two sets of gain functions differ the most. However, even in this region, the differences are not dramatic. Throughout most of the rest of the domain, the differences are considerably smaller. The impact of these differences in the gain functions is most evident in Fig. 2.12 which shows the errors made in the geopotential field during the first 10 sols of assimilation using different gain functions. The dotted line corresponds to using the zeroth iterate Wiener gains. It takes 4 sols to reach a steady-state in which about 30% of the geopotential variance is unexplained. The solid line corresponds to the third iterate Wiener gains derived ignoring topog-

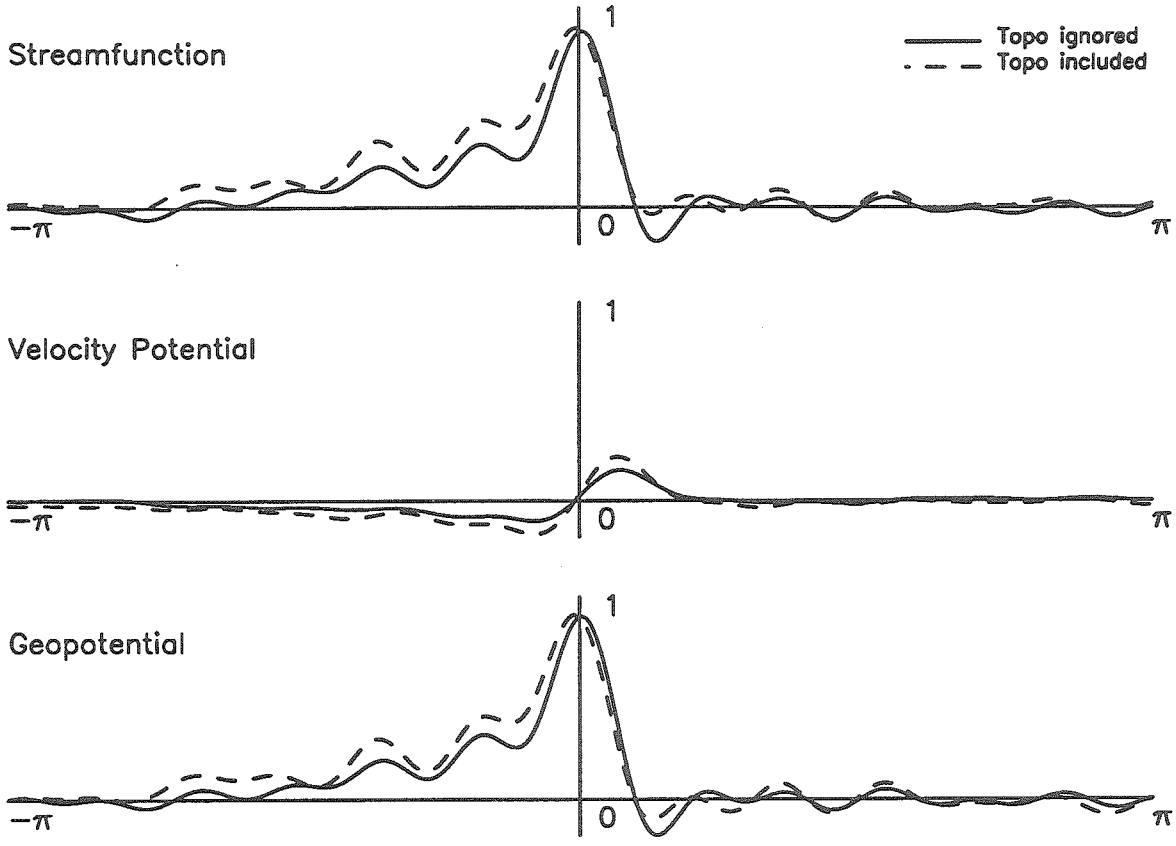


Fig. 2.11: Effect of topography on the gain functions. The dashed line is for the Tharsis region, where the topographic relief is greatest. The gain function for this region differs the most from the gain function derived by ignoring topography (solid line). The differences are not astounding in this example, and are much smaller at other longitudes.

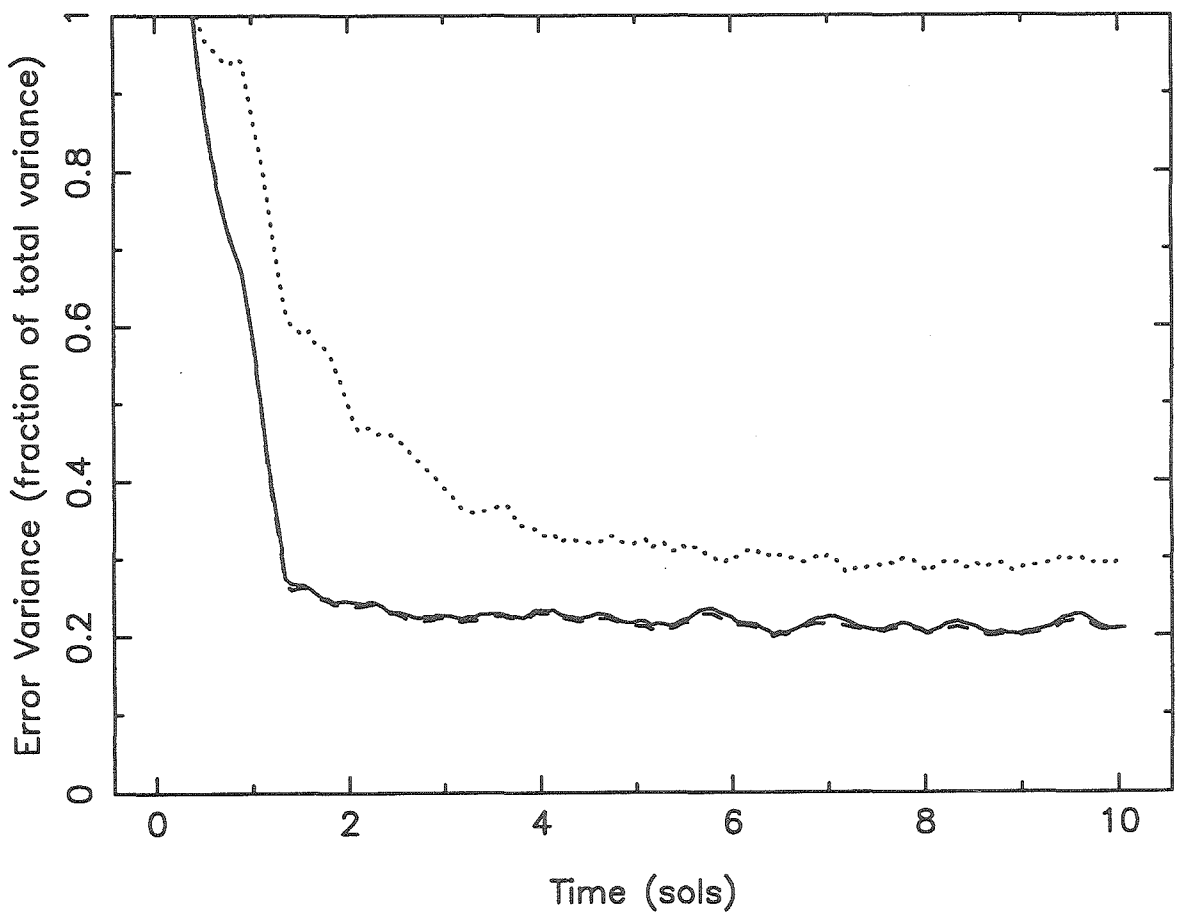


Fig. 2.12: Performance difference in analyzing the geopotential field using gain functions derived ignoring the presence of topography and gain functions derived explicitly including topographic effects, for a 10 sol run. The predictive model was started from a zero state. The solid line corresponds to the gain functions derived ignoring topography, while the dashed line corresponds to the gain functions with topography included. The performance difference is insignificant, particularly when compared to the performance of the zeroth iterate Wiener gains (dotted line).

raphy. The dashed line is the result of using the gain functions including topography. Their performance is essentially indistinguishable from the gain functions derived without including topography. Both reach steady-state in one to two sols, and can explain all but about 22% of the geopotential variance. Similar results were found for the other two state variables, with streamfunction being equivalently well predicted as geopotential ($\sim$ 22% unexplained variance) and velocity potential remaining essentially unpredicted ($\sim$ 88% unexplained variance).

Thus, while local differences can be found between the gain functions that accounted for topography and those that did not, the performance difference is insignificant. The gain functions which included topography outperformed those without topography by less than 1% of explained variance, while the standard symmetric zeroth iterate gain functions performed roughly 8% worse than either of the Wiener gains. Therefore, we believe that it is acceptable to ignore topography and any other satellite longitudinal dependence of the gain functions.

2.7 Spherical Shallow Water Model

Since our first operational application of the Wiener filter will be on a full-scale multi-layer Martian GCM (Pollack *et al.* 1981, 1990, Haberle *et al.* 1992), a high-resolution, single-layer primitive equation model appeared appropriate to serve as the last milestone in our testing hierarchy. We have thus derived the Wiener gains for the single-layer spherical shallow water model discussed in Keppenne (1992). This model is based on the following

dimensionless system of prognostic equations:

$$\begin{aligned}
\frac{\partial}{\partial t}\Delta\Psi &= -\nabla\cdot[\mathbf{u}(\Delta\Psi+f)] - \tau U\Delta\Psi - \gamma\Delta^4\Psi - \epsilon\Delta(\Psi-\Psi^\star) \\
\frac{\partial}{\partial t}\Delta\chi &= \mathbf{k}\cdot\nabla\times[\mathbf{u}(\Delta\Psi+f)] - \tau U\Delta\chi - \gamma\Delta^4\chi - \Delta[\frac{\mathbf{u}\cdot\mathbf{u}}{2}+\Phi] \\
\frac{\partial}{\partial t}\Phi &= -\nabla\cdot[\mathbf{u}(\Phi-\Phi_b)] - (\varphi-\varphi_b)\Delta\chi
\end{aligned} \tag{2.9}$$

The notation and non-dimensionalization are different from those of earlier sections. The first two lines of (2.9) above are the vorticity-divergence form of the momentum equation and the third line of (2.9) is the mass conservation equation. $\Psi(\lambda,\theta,t)$, $\chi(\lambda,\theta,t)$, and $\Phi(\lambda,\theta,t)$, are the streamfunction, velocity potential, and the deviation of the geopotential height of the free surface $h(\lambda,\theta,t)$, representing the top of the atmosphere, from its global average φ, so that $\Phi+\varphi = gh$, where g is the acceleration of gravity. λ is longitude, θ latitude, and t time. The dimensionless time unit is equal to the Martian rotation period $2\pi/\Omega$ and the dimensionless length unit is the equatorial radius of Mars. $\mathbf{u}(\lambda,\theta,t)$ is the horizontal fluid velocity in these units, $f = 2\Omega\sin\theta$ is the Coriolis parameter, τ is a drag coefficient, γ is a diffusion coefficient, U is the mean square-root zonal kinetic energy per unit mass, $\Phi_b(\lambda,\theta)$ is topography and Δ is the Laplacian operator. $\mathbf{k}$ is the local vertical unit vector. $\Psi^\star$ is the forcing field. The forcing-dissipation term, $\epsilon\Delta(\Psi-\Psi^\star)$ is applied only to the zonal part of the streamfunction field.

The global mean geopotential height φ of the free surface is 50. In dimensional units this height is $(50/g)(r\Omega/2\pi)^2$, about 20 km for Mars. At a

latitude of 30° the radius of deformation, L_D is $r\sqrt{50}/2\pi$, which is ~ 3 times greater than in our earlier examples. The drag coefficient τ is 0.1, and the relaxation time, $1/\epsilon$, is 10 sols, which is a factor of 2.5 times greater than in our earlier examples. The diffusion coefficient γ is 5×10^{-12}.

Equation (2.9) is expanded in terms of spherical harmonics and its spectral form is integrated at rhomboidal truncation R15 using an associated Gaussian grid with 40 Gaussian latitudes and 48 meridians. A realistic mountain field is interpolated from the topography of Pollack *et al.*'s (1990) Martian GCM whose horizontal resolution is equivalent to the one used here. The transform method (Orszag 1970, Eliassen *et al.* 1970, Bourke 1972) is used to iterate back and forth between the grid-point and spectral representations of the model variables.

This model's global behavior is fairly representative of the Martian atmosphere, and locally, the statistics of its geopotential time series match well those of the surface pressure data at the sites of Viking Landers I and II.

The modeled spacecraft trajectory was chosen to obey a polar sun-synchronous orbit, with a sampling rate of 1000 geopotential observations per sol - about one observation every 250 km of spacecraft down-track motion. Since Mars Observer would have observed the vertical structure of the temperature field over the entire depth of the Martian atmosphere, one single geopotential observation here amounts to a snapshot of the vertical temperature profile in the case of a multi layer GCM. As in the other experiments, the observations were degraded with 5% Gaussian noise, and the model was randomly forced with red noise so that the 'Truth' and predictive runs were not identical twins. The variance of the random forcing term that affects

the time evolution of each spectral coefficient was set equal to that of the corresponding nonlinear term.

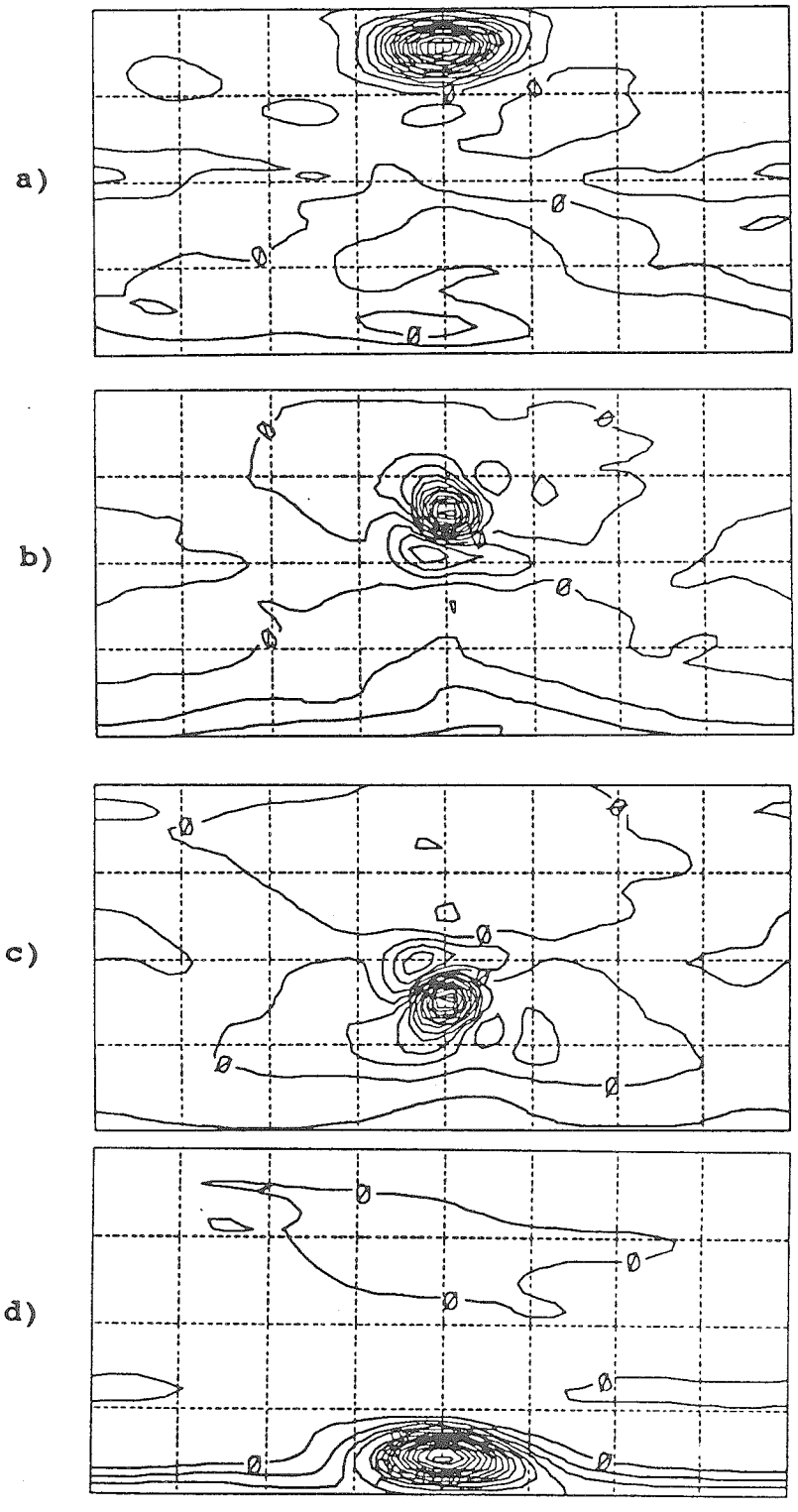
Since this is a nonlinear model, there is no analytic solution to the optimal filtering problem. Although an implementation of the Kalman filter by linearizing the model about its continuously evolving state is conceivable, such an implementation would overwhelm even the fastest current-generation supercomputers. We thus restricted our data assimilation applications on this model to the derivation of the corresponding steady-state Wiener gains. Three iterations were sufficient for the Wiener gains to reach a quasi-steady-state, i.e., a fourth iteration with the predictive model assimilating data using the third iterate gain functions resulted in virtually no change from the third to fourth iterate. Each set of gain functions was computed from the statistics of the second half of a 2000-sol prediction run, continuously forced with geopotential observations of a comparatively long history tape of a 'Truth' run.

In order to successfully initialize the iterative procedure, it was necessary to scale down the zeroth iterate gain functions. When unscaled, these gain functions induced the predictive run's trajectory to oscillate wildly about the 'Truth' and in some cases "blow up." This undesirable behavior arose probably because the weather statistics, which were used in the zeroth iterate Wiener gains, differ more substantially from the steady-state error statistics than in the experiments with the linear models. The result could be an overestimation of the error covariances by the zeroth iterate Wiener gains. It is also possible that the zeroth iterate Wiener gains did not have the proper balance between variables (geopotential and wind, for example), and that the

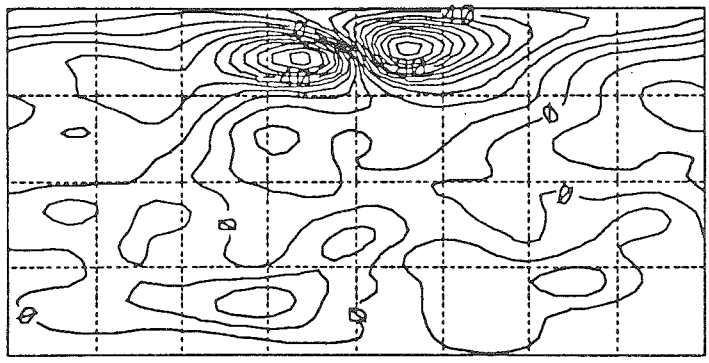
improper balance excited some of the model's unstable modes. Whatever the case, multiplying the zeroth iterate Wiener gains by a number from 0.1 to 0.5 was sufficient to alleviate this problem. The problem did not occur on subsequent iterations, and no further reductions of the gain functions were necessary. Once properly initialized, the iteration converged as before.

Contour maps of the longitudinally independent third iterate Wiener gain functions are shown in Fig. 2.13 in the spacecraft's frame of reference, with the sub-spacecraft grid point located along the center meridian at the appropriate latitude. These maps correspond to the Gaussian latitudes closest to 67.5°N, 22.5°N, 22.5°S and 67.5°S. Presumably (since we could not check this assumption with an analytically derived Kalman filter), they show how geopotential observations should be weighted to optimally update the streamfunction (Fig. 2.13 a-d), velocity potential (Fig. 2.13 e-h) and geopotential field (Fig. 2.13 i-l) to steer the predictive model toward the 'Truth' model.

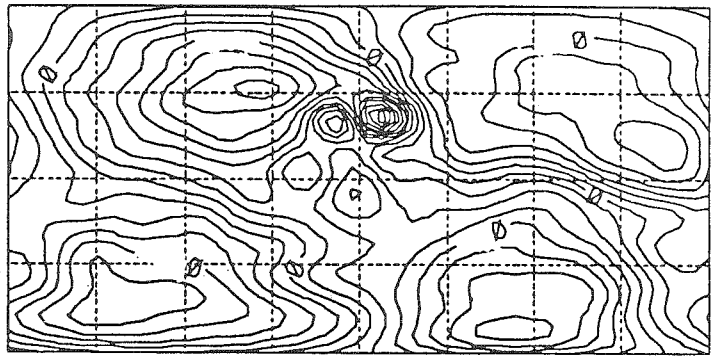
Several points are worth noting about the shapes of these gain functions. First, the longitudinal asymmetry reported in our discussions of the linear models' gain functions is less noticeable, although it is still present (Fig. 2.13 d and l). We are not sure of the cause of this reduction, but suspect that it is due to the greater relaxation time used in this model (10 sols versus 4 sols previously). A longer relaxation time would result in less errors being built up over 1 sol. Thus, differences between the east and west sides of the observation point would also be reduced, and the gains would be more symmetric than in the previous models. Second, there is a substantial departure from geostrophic adjustment, especially in the tropics (see Fig. 2.13 b, c, j, and


a)
b)
c)
d)

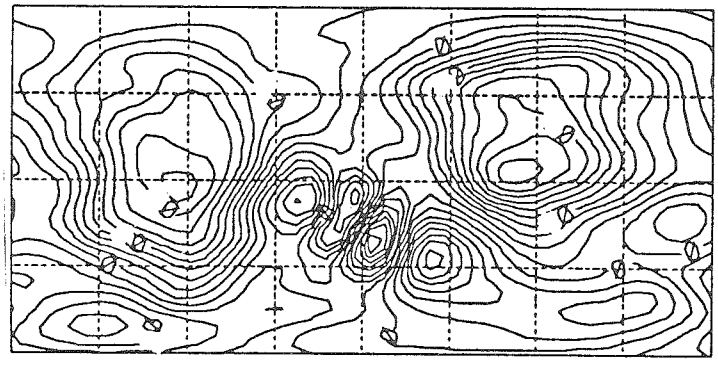
e)



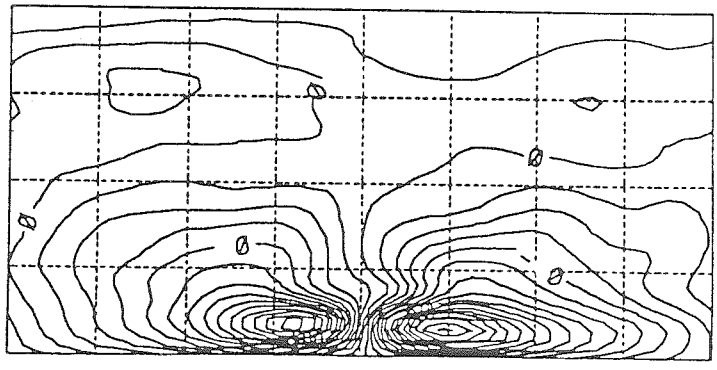
f)



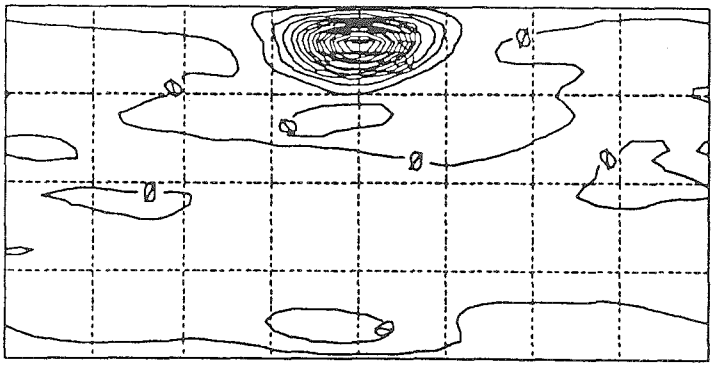
g)



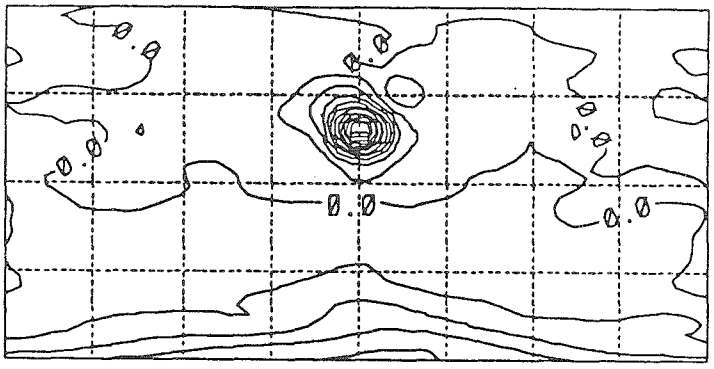
h)



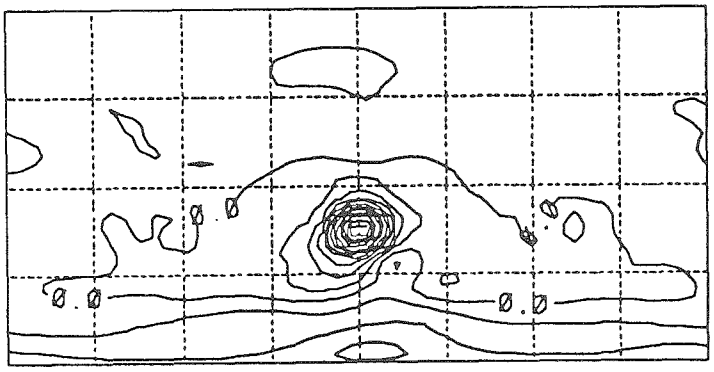
i)



j)



k)



l)

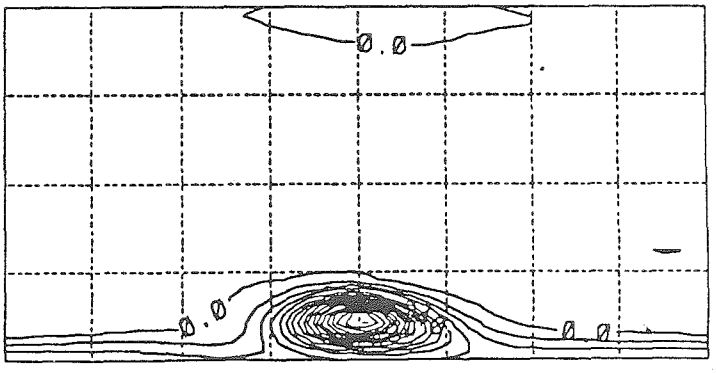


Fig. 2.13: Contour maps showing the longitudinally independent third iterate Wiener gains for the global nonlinear shallow water model in the spacecraft's frame of reference, with the sub-spacecraft grid point located along the center meridian at the appropriate latitude. These maps correspond to the Gaussian latitudes closest to 67.5°N, 22.5°N, 22.5°S and 67.5°S. Fig. 2.13 a-d, e-h, and i-l correspond to the gain functions used to update the streamfunction, velocity potential and geopotential fields, respectively, when observations are taken at these latitudes. These maps show how one should expect the theoretical Kalman gains to be spatially distributed, for this model of Mars' atmosphere when the observational pattern matches that of a single, sun-synchronous polar orbiting satellite.

k) where the Coriolis approximation can not be made. The streamfunction weights are characterized by a dipole pattern at low latitudes (Fig. 2.13 b and c). The geopotential weights (Fig. 2.13 i-l) resemble more the Gaussian bell patterns of operational OI schemes. The velocity potential gain functions (Fig. 2.13 e-h), are of much lesser magnitude than their streamfunction and geopotential counterparts because the velocity potential and geopotential fields are essentially uncorrelated. They display a more complicated organized structure that reflects long-distance correlations in the velocity potential error field.

The analysis performance of the third iterate Wiener gains are compared in Fig. 2.14, which shows the normalized streamfunction, velocity potential, and geopotential error variances for a 20-sol run of the predictive model against a history tape of the 'Truth' model. The predictive model was initialized with a zero initial state. After 20 sols, the third iterate Wiener gains give unexplained variances of 4%, 16% and 3% in the streamfunction, velocity potential and geopotential fields, respectively. Considerably larger residual errors remain when the zeroth iterate weights are used. Note that the streamfunction and geopotential errors decrease exponentially during the first day of the experiment and linearly thereafter, while the velocity potential errors decrease linearly from the start. The magnitude of the errors in assimilating with this model are smaller than those found with the other models tested. This is due to a difference in the strengths of the random forcings as compared to the simpler models. The random forcing completely dominated the behavior of the previous models, while in this model, the zonal flow and topography have a more significant effect on the climatology. Thus,

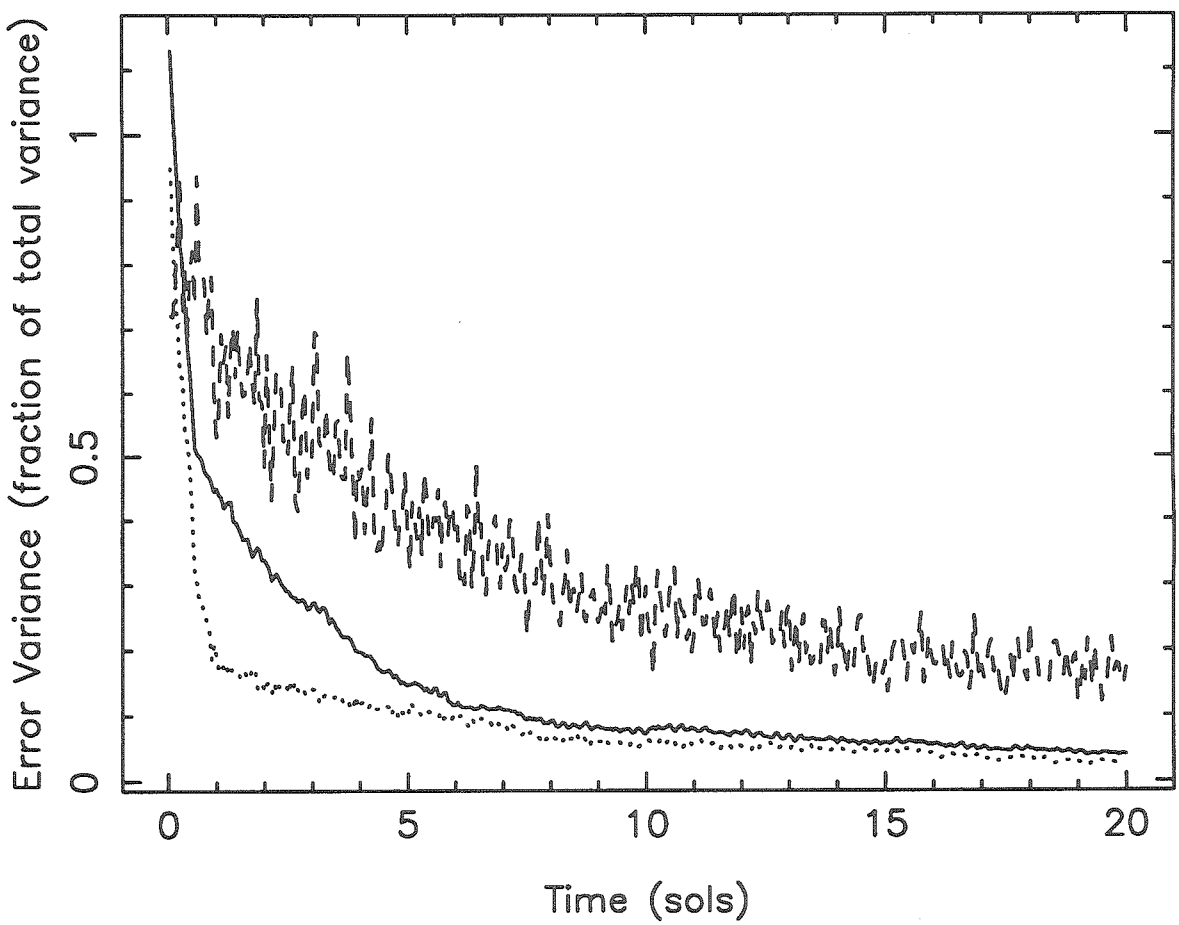


Fig. 2.14: Plot showing the performance of the third iterate Wiener gains derived with the global nonlinear shallow water model. The solid line shows the evolution of the normalized streamfunction error variances during a 20-sol prediction run against a history tape of a 'Truth' run, perturbed by random forcing terms. The dashed and dotted lines show the corresponding evolutions of the velocity potential and geopotential error variances, respectively.

the prediction runs and 'Truth' runs for this experiment were more similar than for the previous models, and analysis performance increased. The main result of the experiments with this model is simply that the assimilation scheme works on this advanced model.

2.8 Conclusion

The main result of this work is the development of a new approach to data assimilation to be used with a single polar-orbiting satellite like Mars Observer. It shares some of the advantages of OI and some of the advantages of Kalman filtering. As in OI the gains are constant in time, but they are equivalent to steady-state optimal Kalman gains. Because the gain functions are computed off-line, this technique has similar computational requirements as OI, and thus considerably less than Kalman filtering. We showed that the effects of longitudinal differences on the shape of the gain functions were not significant. This allows the optimal gain functions to be simplified to be a function of only the observational point's latitude. Furthermore, we showed that for an application like Mars Observer, serial correlations in the observation errors could be safely ignored without a significant loss in predictive performance.

We also demonstrated the feasibility of using these gain functions on a suite of models of increasing complexity, approaching that of a full GCM. These models were observed in only one dynamical variable, as Mars Observer would have done for the Martian atmosphere, yet we were able to make predictions about all three dynamical variables. Furthermore, we veri-

fied the success of an iterative technique in determining these optimal Wiener gains with each model. The next step will be to employ these techniques on a full GCM-scale model, for the assimilation of data from the next Mars atmospheric mission, or perhaps of data from a single polar-orbiting stratospheric-sounding Earth satellite.

2.9 Appendix: Serial and Parallel Assimilation

We show here that serial and parallel assimilation are equivalent when the observation error covariance matrix is diagonal (i.e., uncorrelated observational errors) and there is no change to the state vector due to dynamics (i.e., between time-steps of the model). The proof is by induction: One shows that the equivalence holds for n observations provided it holds for $n-1$ observations. We define $\alpha_{kj}^{(n-1)}$ as the gain matrix, linking the observation at point j to the state vector at point k, in processing $n-1$ observations. Thus, for simultaneous assimilation of $n-1$ observations,

$$\Psi_k^{(n-1)} = \Psi_k^{(0)} + \sum_{j=1}^{n-1} \alpha_{kj}^{(n-1)} (\Psi_j^o - \Psi_j^{(0)}), \tag{2.A1}$$

where $\Psi_k^{(0)}$ is the state vector before the processing begins, and $\Psi_k^{(n-1)}$ is the state vector after $n-1$ observations have been processed. The $\alpha_{kj}^{(n-1)}$ are assumed to satisfy

$$\sum_{j=1}^{n-1} (r_{ij} + v_j \delta_{ij}) \alpha_{kj}^{(n-1)} = r_{ik}. \tag{2.A2}$$

Here $r_{ij} = \overline{\epsilon_i^p \epsilon_j^p}$, where ϵ_i^p and ϵ_j^p are the prediction errors associated with $\Psi_i^{(0)}$ and $\Psi_j^{(0)}$, respectively. Also, $v_j = \sigma_j^2$ is the error variance of the jth observation, and δ_{ij} is the Kronecker delta. Equations (2.A1) and (2.A2) follow from the assumption that serial and simultaneous processing are equivalent for the first $n-1$ observations.

Serial processing can be written

$$\Psi_k^{(n)} = \Psi_k^{(n-1)} + \alpha_{kn}^{(n)}(\Psi_n^o - \Psi_n^{(n-1)}). \tag{2.A3}$$

Note that n denotes both the location of the observation and its order in the processing sequence. The $\alpha_{kn}^{(n)}$ are computed as in (2.3), but the errors ϵ_i^p and ϵ_j^p are the errors in $\Psi_i^{(n-1)}$ and $\Psi_j^{(n-1)}$, respectively. From (2.A1), these errors satisfy

$$\epsilon_k^{(n-1)} = \epsilon_k^{(0)} + \sum_{j=1}^{n-1} \alpha_{kj}^{(n-1)}(\epsilon_j^o - \epsilon_j^{(0)}). \tag{2.A4}$$

We substitute (2.A4) into (2.3), recognizing that the observation errors are uncorrelated with the prediction errors and that $r_{ij} = r_{ji}$. This yields

$$\alpha_{kn}^{(n)} = \frac{r_{kn} - \sum \alpha_{kj}^{(n-1)} r_{nj} - \sum \alpha_{nj}^{(n-1)} r_{jk} + \sum\sum \alpha_{ni}^{(n-1)} \alpha_{kj}^{(n-1)} (r_{ij} + v_j \delta_{ij})}{r_{nn} + v_n - 2\sum \alpha_{nj}^{(n-1)} r_{jn} + \sum\sum \alpha_{ni}^{(n-1)} \alpha_{nj}^{(n-1)} (r_{ij} + v_j \delta_{ij})}. \tag{2.A5}$$

The single sums are over j from 1 to $n-1$. The double sums are over i and j, both of which vary from 1 to $n-1$. With (2.A2) this becomes

$$\alpha_{kn}^{(n)} = \frac{r_{kn} - \sum \alpha_{kj}^{(n-1)} r_{nj}}{r_{nn} + v_n - \sum \alpha_{nj}^{(n-1)} r_{nj}}. \tag{2.A6}$$

Substituting (2.A1) into (2.A3), one can relate $\alpha_{kj}^{(n)}$ to $\alpha_{kj}^{(n-1)}$ for $j < n$:

$$\alpha_{kj}^{(n)} = \alpha_{kj}^{(n-1)} - \alpha_{kn}^{(n)} \alpha_{nj}^{(n-1)}. \tag{2.A7}$$

Equations (2.A6) and (2.A7) follow from (2.A3), the nth step of the serial assimilation. The remainder of the proof is showing that (2.A6) and (2.A7) are consistent with the expression for simultaneous assimilation, the n-dimensional version of (2.A2), i.e., with $(n-1) \rightarrow n$. First separate out the nth row and nth column:

$$(r_{nn} + v_n)\alpha_{kn}^{(n)} + \sum_{j=1}^{n-1} r_{nj}\alpha_{kj}^{(n)} = r_{nk}. \tag{2.A8}$$

$$r_{in}\alpha_{kn}^{(n)} + \sum_{j=1}^{n-1} (r_{ij} + v_j\delta_{ij})\alpha_{kj}^{(n)} = r_{ik}, \tag{2.A9}$$

where i varies from 1 to $n-1$. Then substitute for $\alpha_{kj}^{(n)}$ in (2.A8) and (2.A9), using (2.A7). Equation (2.A9) becomes an identity and (2.A8) becomes (2.A6), which shows that the equations are consistent. In other words, serial and simultaneous assimilation are equivalent for n observations provided they are consistent for $n-1$ observations. Since they are equivalent for $n = 1$, they are equivalent for all n. A corollary is that the result of serial assimilation is independent of the order with which the observations are processed.

2.10 References

BAKER, W.E., S.C. BLOOM, J.S. WOOLLEN, M.S. NESTLER, E. BRIN, T.W. SCHLATTER AND G.W. BRANSTATOR, 1987: Experiments with a three-dimensional statistical objective analysis scheme using FGGE data. *Mon. Wea. Rev.*, **115**, 272-296.

BANFIELD, D., A.D. TOIGO, A.P. INGERSOLL, D.A. PAIGE, 1994: Martian weather correlation length scales from Viking IRTM T15. *in preparation.*

BARNES, J.R., 1980: Time spectral analysis of midlatitude disturbances in the Martian atmosphere. *J. Atmos. Sci.*, **37**, 2002-2015.

BARNES, J.R., 1981: Midlatitude disturbances in the Martian atmosphere: A second Mars year. *J. Atmos. Sci.*, **38**, 225-234.

BENGTSSON, L. AND N. GUSTAVSSON, 1971: An experiment in the assimilation of data in dynamical analysis. *Tellus*, **23**, 328-336.

BERGMAN, K.H., 1979: Multivariate analysis of temperatures and winds using optimum interpolation. *Mon. Wea. Rev.*, **107**, 1423-1444.

BIERMAN, G.J., 1977: *Factorization Methods for Discrete Sequential Estimation.* Academic Press, NY, 244pp.

BOURKE, W., 1972: An efficient, one-level, primitive-equation spectral model. *Mon. Wea. Rev.*, **100**, 683-689.

COHN, S.E. AND D.F. PARRISH, 1991: The behavior of forecast error covariances for a Kalman filter in two dimensions. *Mon. Wea. Rev.*, **119**, 1757-1785.

DALEY, R., 1992a: The effect of serially correlated observation and model error on atmospheric data assimilation. *Mon. Wea. Rev.*, **120**, 164-177.

DALEY, R., 1992b: Estimating model-error covariances for application to atmospheric data assimilation. *Mon. Wea. Rev.*, **120**, 1735-1746.

DIMEGO, G.J., 1988: The National Meteorological Center regional analysis system. *Mon. Wea. Rev.*, **116**, 977-1000.

ELIASSEN, E., B. MACHENHAUER, AND E. RASSMUSEN, 1970: *On a numerical method for integration of the hydrodynamical equations with spectral representation of the horizontal fields* Rept. 2, Institute for Theoretical Meteorology, University of Copenhagen, 35pp.

ESPOSITO, P.B., W.B. BANERDT, G.F. LINDAL, W.L. SJOGREN, M.A. SLADE, B.G. BILLS, D.E. SMITH AND G. BALMINO, 1992: Gravity and topography. In: H.H. Kieffer, B.M. Jakosky, C.W. Snyder, M.S. Matthews (Eds.), *Mars.* The University of Arizona Press, Tucson, pp.209-248.

GELB, A., ED., 1974: *Applied Optimal Estimation.* The M.I.T. Press, Cambridge, p.142-143.

GHIL, M., 1989: Meteorological data assimilation for oceanographers. Part I: description and theoretical framework. *Dyn. Atmos. Oceans*, **13**, 171-218.

GHIL, M., S. COHN, J. TAVANTZIS, K. BUBE AND E. ISAACSON, 1981: Applications of estimation theory to numerical weather prediction. In: L. Bengtsson, M. Ghil and E. Källén (Eds.), *Dynamic Meteorology: Data Assimilation Methods.* Springer-Verlag, New York, pp.139-224.

HABERLE, R.M., J.B. POLLACK, J.R. BARNES, R.W. ZUREK, C.B. LEOVY, J.R. MURPHY, H. LEE, J. SCHAEFFER, 1992: Mars atmospheric dynamics as simulated by the NASA Ames general-circulation model. 1. The zonal-mean circulation. *J. Geophys. Res.*, **98**, 3093-3123.

JAZWINSKI, A.H., 1970: *Stochastic Processes and Filtering Theory.* Academic Press, New York.

KEPPENNE, C.L., 1992: Orographically-forced oscillations in a dynamical model of the Martian atmosphere, *Icarus*, **100**, 598-607.

MCCLEESE, D.J., R.D. HASKINS, J.T. SCHOFIELD, R.W. ZUREK, C.B. LEOVY, D.A. PAIGE, F.W. TAYLOR, 1992: Atmosphere and climate studies of Mars using the Mars Observer pressure modulator infrared radiometer. *J. Geophys. Res.*, **97**, 7735-7757.

MILLER, R.N., 1986: Toward the application of the Kalman filter to regional open ocean modeling. *J. Phys. Oceanogr.*, **16**, 72-86.

ORSZAG, S., 1970: Transform method for the calculation of vector-coupled sums: application to the spectral form of the vorticity equation, *J. Atmos. Sci.*, **27**, 890-895.

PARRISH, D.F. AND J.C. DERBER, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747-1763.

POLLACK, J., C. LEOVY, P. GREIMAN, AND Y. MINTZ, 1981: A Martian general circulation experiment with large topography, *J. Atmos. Sci.*, **38**, 3-29.

POLLACK, J., R. HABERLE, J. SCHAEFFER, AND H. LEE, 1990: Simulation of the general circulation of the Martian atmosphere: I - polar processes, *J. Geophys. Res.*, **95**, 1447-1474.

RUTHERFORD, I.D., 1972: Data assimilation by statistical interpolation of forecast error fields. *J. Atmos. Sci.*, **29**, 809-815.

Chapter 3

Martian Weather Correlation Length Scales from Viking IRTM T15

3.1 Abstract

Spring and fall equinox Viking infrared thermal mapper (IRTM) 15μm channel atmospheric temperature (T15) observations are used to estimate the weather correlation length scale of Mars in the pressure range 0.5-1 mbar. The results are important in providing a benchmark for validating Martian general circulation models (GCMs), in determining the optimal placement of a network of landers, and in guiding data assimilation efforts for orbiters and landers. Observations of atmospheric temperature are used to compute an atmospheric mean state, as a function of time-of-day and latitude, which

is then subtracted from the observations to yield weather temperature residuals. These residuals are correlated with each other to determine (1) the weather temperature correlation length scale ($\sim$ 1500km) as a function of latitude and (2) the globally averaged weather temperature variance ($\sim 4K^2$ for $L_S \sim 0°$, $\sim 11K^2$ for $L_S \sim 180°$). Good general agreement is found in comparing to the Rossby radius of deformation and to inferences made from other data sets. The results are also compared with GCM results yielding generally good agreement, with some interesting differences in the magnitudes of the weather temperature variances. An estimate is made of $\sim$ 110 landers needed to globally observe the weather of Mars, although the southern hemisphere could be observed with $\sim$ 10.

3.2 Introduction

Mars has weather. Baroclinic waves and perhaps even frontal systems have been observed passing over the Viking Landers (e.g., Barnes 1980, Tillman *et al.* 1979). The Mariner and Viking Orbiters observed many types of clouds, which resemble what we consider weather on Earth (Kahn 1984). Local and global dust storms have been observed on Mars for roughly a century (Martin and Zurek 1993). However, although we are aware that there is weather on Mars, we know very little about it. The simplest details of the weather are still poorly constrained by the existing analyses of the Martian atmospheric data sets. For example, the length scale over which a weather system can be considered a coherent body has not been directly determined. The main focus of this work is to quantify the length scale of weather systems on Mars.

The weather correlation length scale is important for a number of reasons. The first is simply to better understand the nature of the weather systems on Mars. The length scale and shape of weather correlations are indicative of the spatial coherence of the atmospheric variations, and perhaps the dominant lengths of waves in the atmosphere. These physical quantities of the Martian atmosphere and its weather add a crucial benchmark with which to evaluate the performance and validity of existing Martian GCM's. While other work has allowed the temporally averaged meridional cross-sections in these models to be well compared with observations (e.g., Santee and Crisp 1993, Haberle *et al.* 1993), weather phenomena in these models have been constrained mainly by the Viking Lander observations located at just two surface points (Barnes *et al.* 1993).

Another reason for wanting to know the weather correlation length scale is for use in the assimilation of observational data into numerical models: data assimilation. With Mars Observer, it was planned to take the atmospheric data from the pressure modulator infrared radiometer (PMIRR) instrument which would have given continuous soundings of the temperature structure of Mars' atmosphere and assimilate them into a Martian GCM (Banfield *et al.* 1994). This can be realized with possible future Mars Observer rebuilds, which could yield an accurate estimate of the complete Martian atmospheric state at all times during the mission. However, to properly insert data into a model, the length scale over which weather systems are correlated is needed. With an optimally designed assimilation scheme, an observation at one location will influence the state of the model over a region specified roughly by this weather correlation length scale. If this length scale is not properly

specified, significant performance decreases are found in the data assimilation system's accuracy (Banfield *et al.* 1994). The technique outlined in Banfield *et al.* (1994) requires using a GCM to predict the weather correlation length scales. Determining this length scale from data will allow us to better understand the errors made in using a GCM as a substitute for Mars.

Seaman (1977) showed that the optimal placement of a network of landers is directly related to the weather correlation length scale. Thus, understanding the nature of this quantity, as it changes with latitude, season and dustiness of the atmosphere, can help us intelligently decide where to place landers for future meteorological investigations of Mars, such as a meteorological network.

In this work, we determine the Martian weather correlation length scale using Viking Orbiter IRTM T15 observations. However, before proceeding to that portion of the work, we present the results of some techniques used to estimate the order of magnitude of this quantity. We do this using simple theory and Viking Lander meteorological data. Then we proceed to discuss the techniques we use to extract only weather effects from the IRTM T15 data, and produce the weather correlations. After presenting the results, we compare them with the simple estimates, results from the Earth, and published results from a Mars GCM and other analyses of observations.

3.3 First Estimates

3.3.1 Theory

Some simple estimates of the weather correlation length scale can be inferred from theory. Using quasi-geostrophic theory, we expect the weather correlation length scale to be roughly the same as the Rossby radius of deformation, the length scale at which buoyancy and rotational effects are comparable (e.g., Ghil *et al.* 1979, Krauss *et al.* 1990). The Rossby radius of deformation can be defined by $L_D = NH/f$ where N is the buoyancy frequency of the atmosphere, H is the equivalent depth, and f is the coriolis parameter. For Mars, N is about 0.01 hz, H is about a scale height, 11.5 km (Barnes 1984), and $f = 2\Omega \sin\phi$, yielding $L_D \sim 800\text{km} \times (\sin\phi)^{-1}$. Thus for mid-latitudes, where the atmospheric wave activity is probably centered, we might expect a weather correlation length scale of order 1000 km.

3.3.2 Viking Lander Data

There were two Viking Landers operating on the surface of Mars from September 1976 until April 1980. Viking Lander 1 continued to return data until November of 1982. Among the data returned by the landers were surface pressure, temperature and wind observations at the two sites. Barnes (1980, 1981) interpreted these data as indicative of baroclinic waves propagating in the Martian atmosphere. He reported frequency spectra as a function of season for the pressure, wind and temperature data taken at the site of Viking Lander 2 (48°N, 226°W).

From his results, one can surmise the correlation time of Martian weather, at least at the latitude and times the observations were taken. The widths of the peaks in the frequency spectra are indicative of the inverse of the correlation times of the atmospheric waves. Reading from Figs. 12, 13 and 14 in Barnes (1980), a rough number of $0.08\,\mathrm{sol}^{-1}$ for the full width at half maximum is found from the spectra during fall, winter, and spring. If we assume these spectral peaks represent waves which have oscillatory and damped components of the form: $e^{i\omega t - t/\tau}$, and then Fourier transform this to yield the resulting power spectral density, we find:

$$P(\omega) = \frac{1/\tau}{1/\tau^2 + \omega^2}. \tag{3.1}$$

Equating the width of this power spectral density with that found from Barnes (1980), we find a coherence time of $\tau \sim (\pi \Delta f_{fwhm})^{-1} \sim 4$ sols. While this is not the final quantity that we are seeking, it is an interesting result along the way to deriving the correlation length scale. The coherence time of Earth's weather is somewhat shorter than this, as can be similarly detected looking at surface pressure power spectra for the Earth. Hamilton and Garcia (1986) publish frequency spectra for surface pressure at several sites on the Earth, which generally exhibit a structure which is more like "white noise" than those of Mars. However, the higher time resolution spectra in Hamilton and Garcia's Figs. 11 and 12 show a characteristic width of about $0.0055\mathrm{hour}^{-1}$ for an equatorial normal mode of the atmosphere. Thus, these spectra correspond to a correlation time of less than 60 hours, or 2.5 days. As an aside, the greater correlation time-scale for disturbances in the

Martian atmosphere may make the Martian atmosphere more predictable than that of the Earth.

We can use the correlation time to estimate the correlation length scale one should find at Mars. Given the phase speeds of the waves on Mars and their correlation times, one can calculate the distances over which the waves would travel while they remain coherent disturbances. This distance is not precisely the same as their instantaneous spatial correlation length, but should be similar to it. By assuming geostrophy between the wind and pressure observations, Barnes (1980) also computed phase speeds for the waves, with values between 5 and 15 m s^{-1} at the location of Lander 2. Thus, with correlation times of order 4 sols, a correlation length scale of order 3000 km is expected at this latitude.

Another not entirely unique approach for estimating the correlation length scale can be employed for the second year of Mars Viking Lander data (Barnes 1981). Barnes (1981) again assumed geostrophy between the wind and pressure measurements. From the cross-spectra of the observations, he derived wavenumber frequency spectra for the waves travelling around Mars. He states that the reported variances in the derived zonal and meridional wavenumbers are indicative of the spread of the pressure wavenumber spectrum at a given frequency. Thus, we can use the variances to estimate the width of the wavenumber spectra of these Martian disturbances. Then these widths can be used to infer a correlation length, just as the frequency widths are indicative of a correlation time. The values reported by Barnes are a standard deviation in the zonal wavenumber spectra of about 1 wavenumber at 48°N for fall and winter. If that standard deviation is taken to be the

half width at half maximum of a spectral peak, then a correlation length of about 2000 km is found in the zonal direction at this latitude. Barnes also reports meridional wavenumbers and standard deviations, which we use to infer a meridional correlation length scale. The meridional wavenumber spectra have standard deviations of about 2.5 wavenumbers. This value suggests a correlation length scale of about 1000 km in the meridional direction at this latitude. This result interestingly suggests a possible difference in the length scale depending on the direction in which the correlations are performed. That is, this result suggests that weather is correlated over longer distances in the zonal direction than in the meridional direction at this latitude. Nevertheless, these two values are similar to the value derived using the preceding technique. This is in part a reflection of the fact that both techniques essentially make the same assumptions (geostrophy) in handling the data sets to arrive at the same quantity (the length scale). However, the results are individually interesting as they are from two separate years at Mars.

3.4 Viking IRTM Data

The best method of determining the weather correlation length scale is to take observations of the whole planet at one time and correlate the weather perturbations with each other. While highly instructive, the Viking Lander data were hardly a *global* data set. The Viking Orbiters had an infra-red instrument (IRTM) which was able to sense the atmospheric temperature at a pressure of about 0.6 mbar (about 25 km elevation) in the atmosphere using

its $15\mu m$ channel (Martin and Kieffer 1979). This data set is the focus of this body of work, and we have used it to determine the weather temperature correlation length scale over most of Mars for two seasons. This data set is most like the scenario described above, that of a global simultaneous view of the true atmospheric state. While the observations are not both global and simultaneous, there are observations covering nearly half of the globe within an hour. For the weather systems on Mars, observations within an hour of each other should be effectively the same as simultaneous observations. It is, however, also important to note that the data set from this instrument probes a different region of the atmosphere than that studied by the Viking Landers.

We chose two subsets of the IRTM data to examine in detail, namely 16 weeks centered around $L_S \sim 0°$ (northern spring equinox) and 8 weeks centered about $L_S \sim 180°$ (northern fall equinox). These subsets were chosen because they had the best global coverage of the whole data set, and were obtained during times of low dustiness in the atmosphere. Furthermore, seasonal differences between the fall and spring equinoxes might be evident in the analyses. Eventually, data taken during the winter and summer seasons and during dust storms should also be examined in this manner, but was beyond the scope of this work.

3.4.1 Climatology

The difficulty of using the IRTM data set lies in separating the climatological mean state from the transient weather temperature residuals. In this

case, we are using the word climatology to include all of the systematic temperature effects in the atmosphere: seasonal trends, vertical temperature gradients, and the steady daily temperature fluctuations. This climatology will be primarily a function of latitude and time-of-day, reflecting the strong hemispheric and diurnal effects. However, other effects such as emission angle, local topographic elevation and perhaps longitude of the observation can contribute to the systematic variation of the mean temperature observed in the atmosphere. Seasonal effects will present longer-term variations in the climatology, particularly in the latitudinal distribution of mean temperature, and must be considered as well. In this section, we will discuss the details of our fit to the climatology of the IRTM data set, with which we have attempted to explain as much systematic variance as possible. It is this climatological mean state that we must remove from the observations to yield the weather temperature effects we are ultimately interested in.

As mentioned above, we expect the atmospheric climatology to be mostly described by functions of latitude and time-of-day. Therefore, the primary component of our fit will consist of spherical harmonics in latitude and time-of-day. However, we must also consider the emission angle of the observations, as observations taken at higher emission angles actually sample the temperature of the atmosphere at higher altitudes. We will model this effect by fitting constant vertical gradients in temperature centered about the mean altitude of the observations. Furthermore, we will allow these vertical gradients to be a function of latitude, but not time-of-day. This assumption is suggested from the work of Santee and Crisp (1993) (and Santee, private communication, 1992) who inverted Mariner 9 infrared spectroscopy (IRIS)

spectra to yield temperature profiles of the Martian atmosphere. Their results, from one of the same seasons that we examined ($L_S \sim 0°$), suggest that the vertical temperature gradient in the region sampled by the IRTM instrument is essentially constant with altitude and only a function of latitude. We also attempted to fit spherical harmonics in latitude and longitude to the data, to allow for the possibility of topographically fixed waves in the atmosphere. However, that functionality yielded no consistent results, and was not used in the final analysis. Mathematically, we express our fit in the following manner:

$$\begin{aligned} T_{resid} &= T_{observ} - T^{\phi hz}_{mean} \\ T^{\phi hz}_{mean} &= \sum_{l=0}^{L_{max}} \sum_{m=-l}^{l} T^{\phi h}_{lm} Y_{lm}(\text{latitude, hour}) + (\mu - \overline{\mu}) \sum_{l=0}^{L_{max}} T^{\phi}_{l0} Y_{l0}(\text{latitude}) \end{aligned} \tag{3.2}$$

where μ = cos (emission angle) and $\overline{\mu}$ is the average of that quantity over the period being fit. The Y_{lm}'s are the spherical harmonic functions. The $T^{\phi h}_{lm}$ are the fit coefficients of the spherical harmonics in latitude and time-of-day. The T^{ϕ}_{l0} are the fit coefficients in latitude and emission angle.

To maintain quality in the data and the fits, filtering was done on IRTM's quality flag, emission angle, and local topography of the observations. Only observations tagged with an IRTM quality flag less than 4096 were considered. This represents data with the only potential flaw being unremoved spikes (see Kieffer 1989, PDS Viking IRTM CD-ROM, IRTM.TXT). The emission angle was limited to below 65°, effectively limiting the vertical ex-

tent sampled by the observations. Above this, the constancy of the vertical derivative became suspect. As the 15μm band's weighting function is fairly broad, observations taken over topographically high regions are also suspect. For nadir observations, the weighting function is at 25% of its maximum value at 12 km altitude (Kieffer *et al.* 1976). Observations taken over locations topographically higher than this will clearly be significantly contaminated by surface temperatures, and thus need to be discarded. We chose 12 km as the cut-off, because only a very small portion of the planet is then excluded (Tharsis volcanoes and Olympus Mons) and the IRTM instrument's 15μm band weighting function drops off very quickly below that altitude.

One also quickly recognizes that the climatology will be a function of season, so we limited the time domain covered in each of our climatology fits. The trade-off is that seasonal trends slowly change the mean temperature at any particular location on the planet, but computing a mean from too short a data set causes insufficient coverage, creating large errors in the fit. We could not perform one climatology fit to the entire 8 or 16 week subset of the data analyzed for $L_S \sim 0°$ or $L_S \sim 180°$, as the errors due to seasonal trends would overwhelm the weather temperature fluctuations. We found that a practical balance of 14 sols (Martian solar days) allowed sufficient coverage, yet negligible seasonal changes and the error in our climatology was similar for both of these effects. Specifically, the seasonal changes were less than or on the order of 1K at any particular location over 14 sols, while the coverage in that time allowed our fits to be accurate to of order 1K as well. The value we quote of about 1K accuracy in the fits is found from comparing 2 similarly good fits, e.g., an $L_{max} = 7$ fit and an $L_{max} = 6$ fit. Their differences in the

regions in which there are data are less than about 1K. Thus, we fit to the 14 sol periods, and then after subtracting off the climatology, combined four or eight 14 sol periods to comprise the larger subsets analyzed. Figure 3.1 shows the coverage allowed in one of the 14 sol periods of the IRTM data that we have analyzed. The number of observations in the 0.34H (1H is defined as 1/24th of a sol) by 3.6° latitude bins range from 0 (white) to 200(black).

Figure 3.2 shows the climatological mean state at a few particular latitudes for one 14 sol period. The figure also shows some of the raw data from which the climatology was derived. At first glance, this fit does not appear good, but keep in mind that the weather temperature fluctuations are present in the data, but not in the climatology fits. We will show later how the weather temperature fluctuations are of order $5K^2$, while the errors in the climatology fits and the observational errors are of order $1K^2$. Thus, this does in fact show good agreement between the data and the climatology in the regions in which there are data.

We computed fits out to $L_{max} = 15$, yet found that $L_{max} = 5$ was always sufficient. That is, the terms which most decreased the χ^2 of the fit were all $L_{max} = 5$. Figure 3.3 shows an $L_{max} = 7$ fit in latitude and time-of-day to a 14 sol period of the data at $L_S \sim 0°$. Most noticeable in this fit is the diurnal component of the daily temperature oscillation. It reaches a peak at about 13H, or just after local noon. The amplitude of this diurnal tide is roughly 14K at the equator, although it is difficult to separate the diurnal tide from the semidiurnal tide in this representation. The semidiurnal component is also noticeable, although the large cold area in the atmosphere at 22H around the equator is mainly due to a hole in the coverage

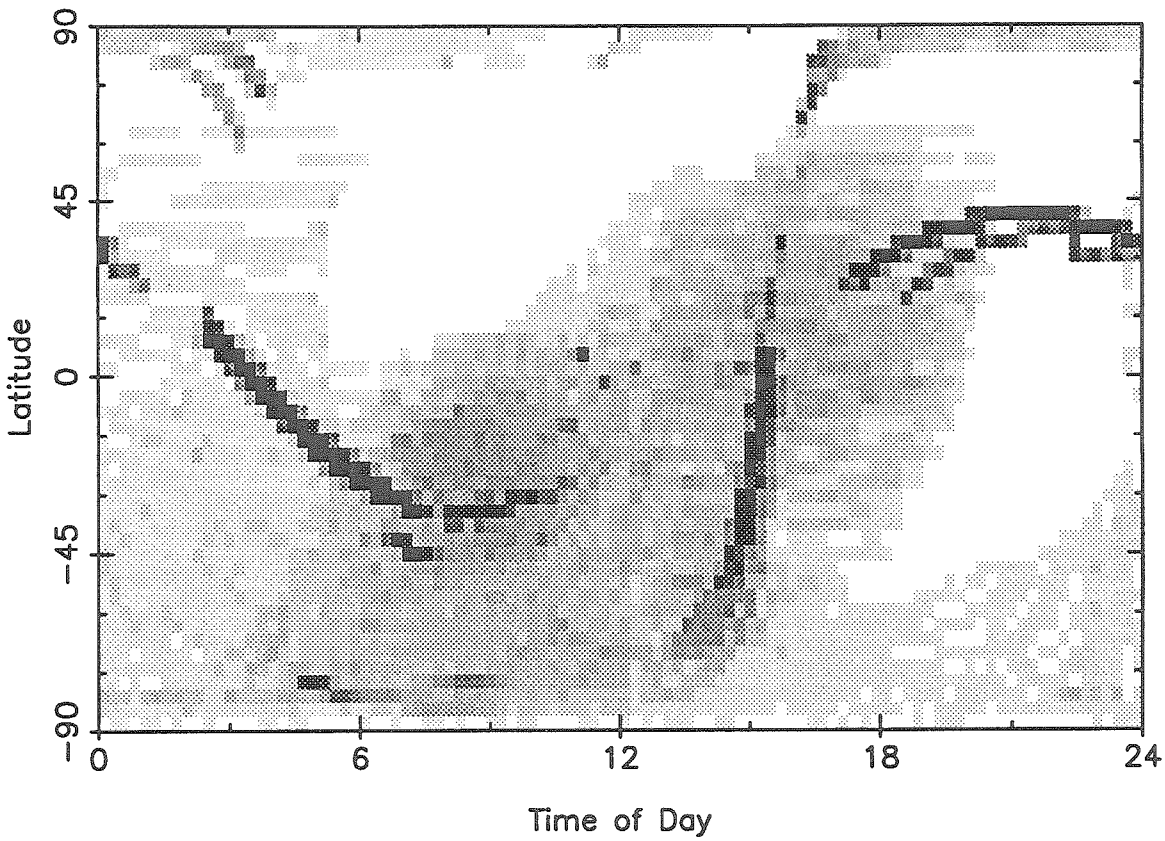


Fig. 3.1: The coverage afforded by Viking IRTM observations from JD 3428 - 3442 (Northern spring equinox, $L_S \sim 0°$). The number of observations in this period is binned by 0.24H and 3.6° latitude, with values ranging from 0 to 484 per bin. The grey scale is adjusted so that white represents a data gap (no observations) and it saturates to black at 200 observations per bin. Note the poor coverage in the North, and at certain times-of-day farther South.

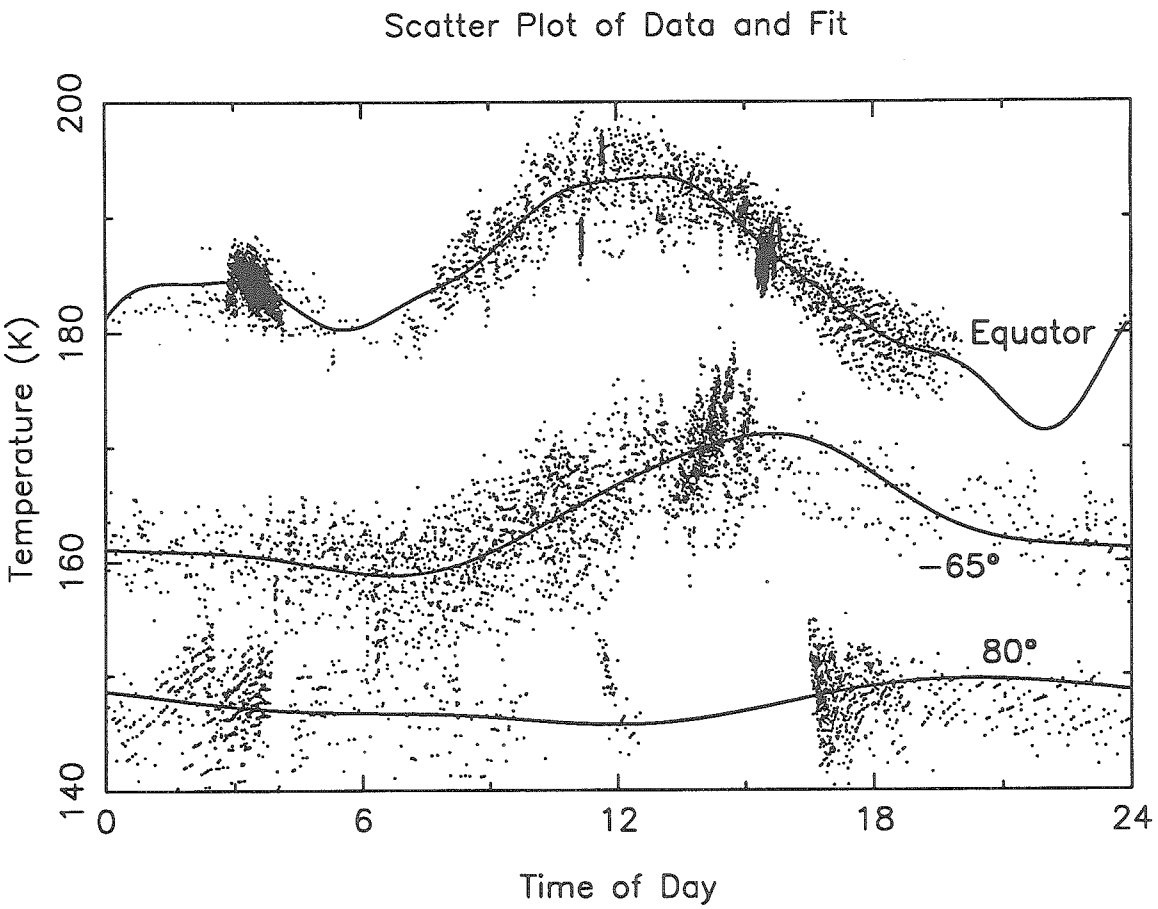


Fig. 3.2: A comparison of the climatology fit and the data used to derive it for one 14 sol period (JD 3428 - 3442) of the Viking IRTM T15 data. Shown on this plot are cross-sections of the global fit at three latitudes: the equator, 65°S, and 80°N. Note the good accord between the data and the fits in regions in which there are data. The discrepancies between the fit and the data mainly represent the weather temperature variations we are primarily interested in.

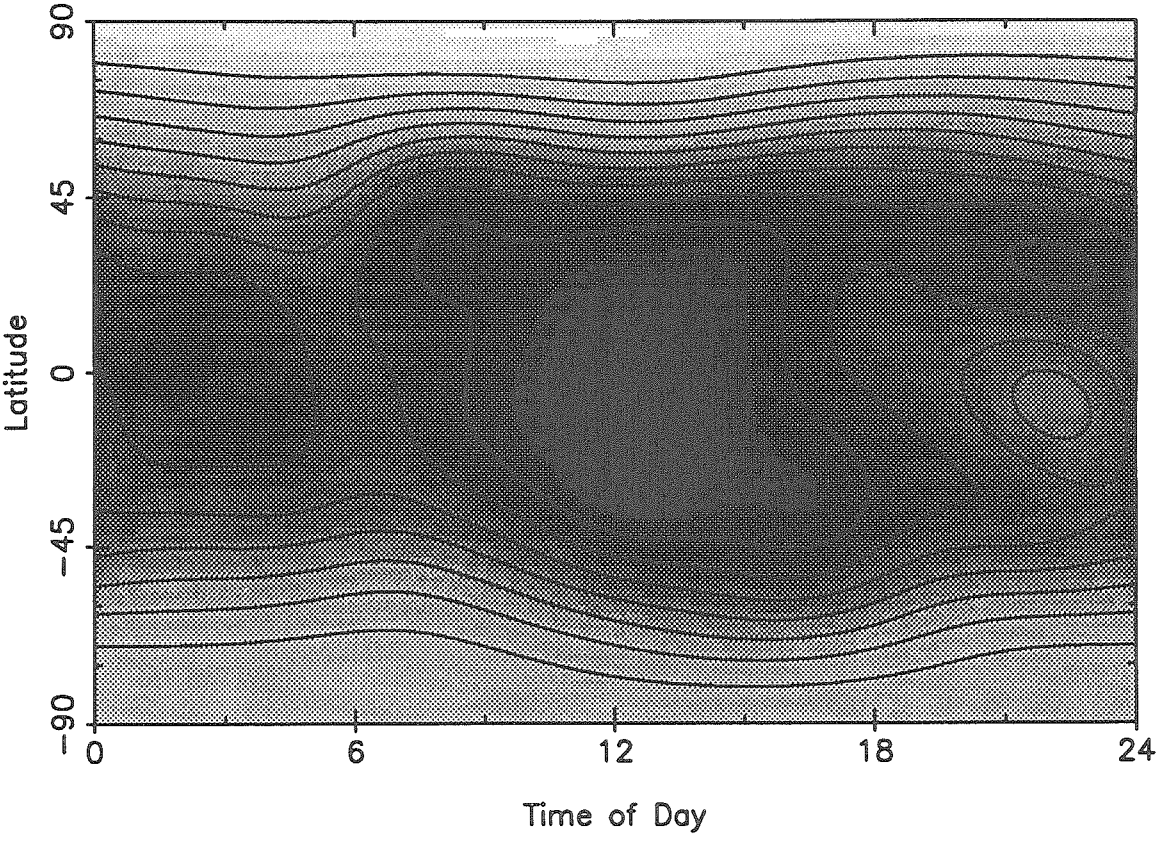


Fig. 3.3: The mean temperature of a 14 sol period of the IRTM data as a function of latitude and time-of-day. Dark represents hot, and the contours are between 149K and 195.4K, spaced by about 4.6K. This fit is in spherical harmonics out to $L_{max} = 7$. The diurnal variation has an obvious peak near the equator at a time-of-day just slightly after local noon. The rough latitudinal symmetry is indicative of this period being near equinox ($L_S \sim 0°$). The local temperature maximum just north of the equator and 22H is spurious and is due to no coverage in that region (see Fig. 3.1).

of the data (see Fig. 3.1). The atmospheric cooling towards the poles is also prominent in this plot, with a difference of 50K between the highest equatorial temperature and the coldest polar temperature. Again, this fit (and thus, the tidal amplitudes) is roughly within 1K of the actual climatology of Mars' atmosphere at this altitude and season in regions which were well sampled by the IRTM instrument. Clearly there is much information in these fits about the structure, amplitude and phase of the thermal tides in the Martian atmosphere at this altitude. However, a detailed analysis is beyond the scope of this paper, but will be explored in later work.

Most of the systematic variation in the mean temperature of the atmosphere is the result of latitudinal gradients and daily temperature fluctuations. Figure 3.3 represents these functionalities. However, recall that we also fit functions describing the vertical derivatives of the temperature profiles as a function of latitude and emission angle of the observation. In addition to daily fluctuations, latitudinal and vertical gradients of the mean atmospheric temperature, one might expect to find stationary or topographically-fixed temperature fluctuations in the atmosphere as well. These would be manifest as systematic temperature deviations as a function of location on the planet, or latitude and longitude. As mentioned above, we did not fit terms as a function of latitude and longitude to the data because they showed little consistency between consecutive 14 sol periods. That is, the differences between the latitude and longitude fits of one 14 sol period and the next were of the same order as the power in those functions themselves. As a comparison, the differences between latitude and time-of-day fits of one 14 sol period and the next were less than about 10% of the power in the latitude and time-of-

day functions. Furthermore, the power in the latitude and longitude terms was roughly only 10% of the power in the latitude and time-of-day terms. Finally, visually the latitude and time-of-day fits were very consistent from one 14 sol period to the next, while the latitude and longitude fits bore no resemblance from one to the next. Thus, either topographically fixed waves changed significantly on a time scale of 14 sols, or simply could not be discerned from this data set. We estimate that a minimum amplitude of order 5K should have been observable in latitude and longitude variations, as this is the apparent level of noise present in the terms with this functionality.

After subtracting the climatological fits from the data, scatter plots of the residuals were used to check for any systematic variation left as a function of any of the independent variables, i.e., emission angle, latitude, longitude, local topography, etc.. No apparent correlations were found in any of the independent variables. Thus, the combination of the data filtering and the functionality of our fits adequately describes the systematic variation in the data set.

3.4.2 Weather Correlations

Having derived the climatological mean state, we then subtract it from the raw data to leave the transient weather temperature fluctuations. In this section, we will discuss how we compute the weather temperature correlations and their variances as well as present the results of these calculations.

For each datum, its departure from local climatology is now the part of the data set we are interested in. We correlated each weather temperature

residual with each of the other nearly contemporaneous (within 1 Earth hour, i.e., $\sim$ 1 Mars hour) weather temperature residuals in its 20° latitude bin. We binned the variances and covariances by the latitude of the observation and the separation distance between the two observations being correlated.

$$\rho(\phi, d) = \frac{\mathrm{cov}[T_{\mathrm{resid}}(\phi_1, \theta_1), T_{\mathrm{resid}}(\phi_2, \theta_2)]}{\mathrm{var}[T_{\mathrm{resid}}(\phi)]} \tag{3.3}$$

where ρ represents the correlation, and ϕ is the latitude bin of the observations. d is the two observations' separation distance. ϕ_1, θ_1 and ϕ_2, θ_2 represent the latitude and longitude of the first and second observations respectively. Attempts were made to also compute the correlations as a function of the azimuth angle from north between the two observations being correlated, but poor results were found and will not be reported here.

We had to choose a time window in which we could consider observations to be simultaneous, at least from the perspective of trying to correlate moving weather systems. As mentioned previously, we can expect a phase speed of these systems of roughly 20 m s^{-1} (at least in the mid-latitudes) (Barnes 1981). Our previous simple estimates of the weather correlation length scale suggest a length scale of order 2000 km. We can consider two observations of a weather system simultaneous if it has moved only a small fraction of its own size in the time between the observations. So as not to automatically select out the same length scale as previously estimated we will use conservative estimates to determine an acceptable time window. If we require a 500 km disturbance to move less than one fifth of its size, it could only move 100 km. If it is traveling at 30 m/s, this means it could go for about 1 hour before

it propagated 100 km. Thus, we use a window of 1 hour during which we consider observations to be simultaneous. We also tried performing all the calculations with a time window of 3 hours, which negligibly changed the results.

We defined above how we calculated the correlations as a function of latitude and separation distance between the points being correlated. However, the weather temperature variance in the denominator of the correlation is not simply the average product of the observational residuals with themselves, because such a product would include observational noise and errors in the climatology. That is, the total residual variance will be the sum of the weather temperature variance (the part we're interested in), the observational error variance and the climatological fit error variance.

$$\sigma^2_{\text{total}} = \sigma^2_{\text{wx}} + \sigma^2_{\text{obs}} + \sigma^2_{\text{fit}} \tag{3.4}$$

The observational noise variance is not small ($\sim 3K^2$), and must be considered. It can be assumed to be uncorrelated from one observation to the next, as it is primarily due to detector noise (Chase *et al.* 1978), but it will certainly correlate with itself. It is a function of the temperature of the observation as the detector sensitivity changes with observed temperature. The climatological fit error variance is somewhat smaller than the observational error variance ($\sim 0.5K^2$), but is also a function of latitude, and hard to estimate well. We could model these error variances and remove them from the total variance calculations. However, we chose not to do this, as our ability to model these error variances is poor and would likely introduce significant

errors into our weather temperature variance estimates.

To correct the total variances to the weather temperature variances, we use the idea the the weather correlations should approach 1.0 at zero separation. Unless there are significant small scale structures in the weather, the weather correlations should smoothly rise to unity at the origin. Thus, by requiring this of the correlations, we can implicitly remove the observational and climatological fit error variances without modelling them. We have performed this by extrapolating the original correlations at 25 and 75 km separation to zero separation. Then the weather temperature variances were adjusted so that this correlation at zero separation was 1.0. We did attempt to quantify and remove the unwanted variances in the total residual variances, and found results quite similar to that found by extrapolating to zero separation. Nevertheless, we believe that the extrapolation technique introduced less errors in the results.

Although the errors in the climatology fit are small compared to the variance of the weather temperature residuals, there was concern that our correlation length scales might simply represent the smallest length scales in the fits. We tested for this by computing the correlations using climatology fits of differing length scales, namely $L_{max}=2$, $L_{max}=3$, $L_{max}=4$, $L_{max}=5$, $L_{max}=6$, $L_{max}=7$ and $L_{max}=10$. The results were that the correlations of the $L_{max}=5$, $L_{max}=6$, $L_{max}=7$ and $L_{max}=10$ fits were nearly identical (as were the largest terms in the fits), but the $L_{max}=2$, $L_{max}=3$ and $L_{max}=4$ fits differed greatly from the others. These correlations are shown for a particular latitude band and 14 sol period in Fig. 3.4. Thus, we conclude that our correlations are measuring the weather temperature correlation length scale.

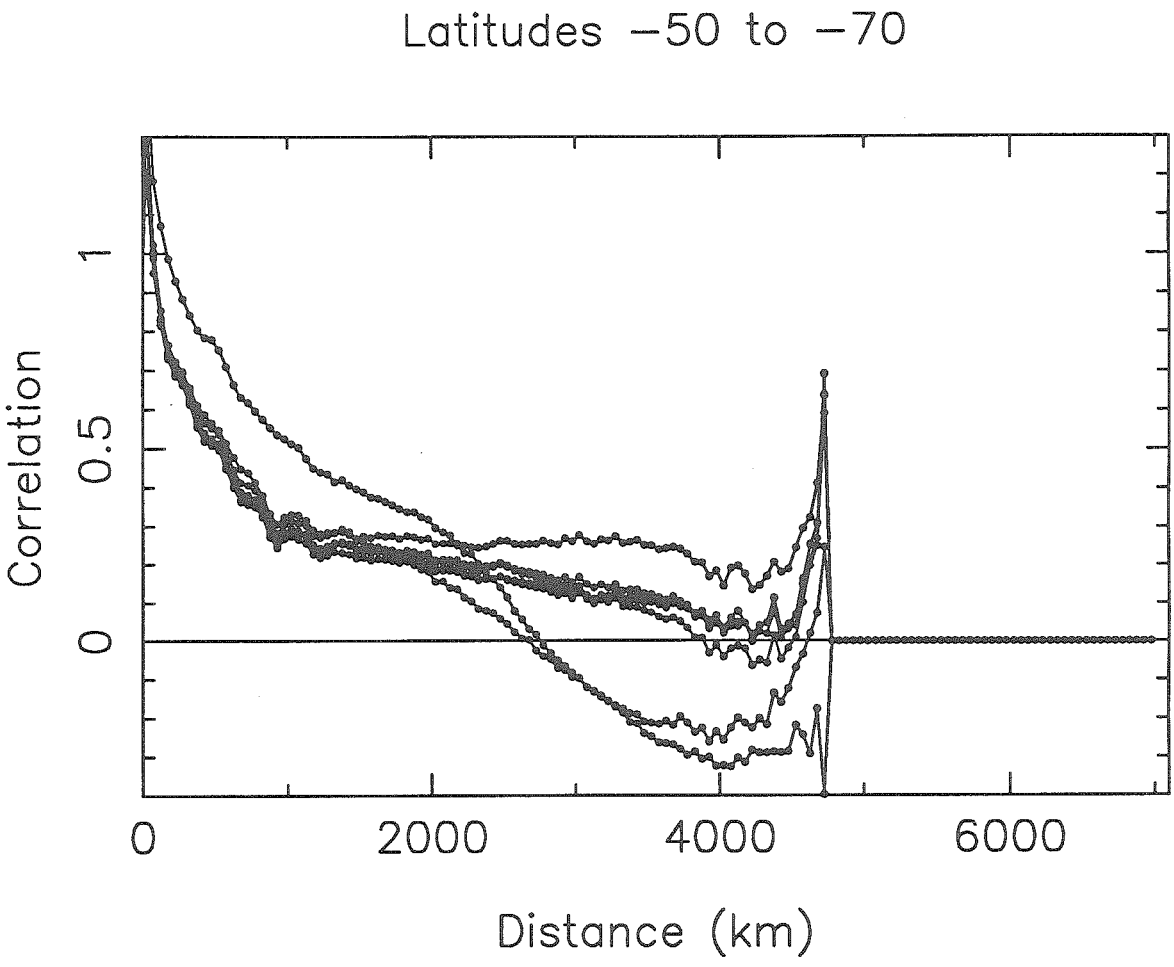


Fig. 3.4: Correlations as a function of distance for a particular latitude band and 14 sol period of the Viking IRTM T15 data. The seven curves on the plot represent different spherical harmonic truncations in the model of the climatology. The four curves which are nearly super-imposed on one another are for $L_{max}=10$, $L_{max}=7$, $L_{max}=6$ and $L_{max}=5$, suggesting that their climatology fits all adequately describe the climatological variations. The other three curves which all differ are for $L_{max}=4$, $L_{max}=3$ and $L_{max}=2$ suggesting that their climatology fits have insufficient resolution to adequately describe the climatological variations.

This also indicates again that the climatology fits can be adequately described by a fit with $L_{max}=5$.

As mentioned before, we computed the climatology fits and correlations of the weather temperature residuals for each 14 sol period of the IRTM data set analyzed. However to reduce noise, we co-added four such periods of 14 sols each (8 weeks) to compute one set of correlations. We analyzed two such 8 week sets centered about $L_S \sim 0°$ to assess the reproducibility of our correlation calculations. That is, for $L_S \sim 0°$, we computed two groups of correlations, one from the 1st, 3rd, 5th and 7th 2 week periods, and the other composed of the 2nd, 4th, 6th and 8th 2 week periods. The differences between the two groups should be indicative of the errors present in the correlations. The 8 weeks of $L_S \sim 180°$ observations were also analyzed in 14 sol periods, but were then all combined to make only one group of correlations to compare with the opposite season.

As an example of the correlations computed, in Fig. 3.5 we present the results for the latitude band centered at 20°N at $L_S \sim 0°$. This latitude band encompasses the Viking Lander 1 site, and thus is instructive for comparisons with the Lander results. The figure shows the correlations from both the first and second group of observations analyzed at this season. The differences between these curves are indicative of the reproducibility of the results, and thus the errors in the correlation curves. While the curves are not identical, they do follow a similar trend out to about 2000 km. Beyond there, the curves are less similar, primarily due to the fact that fewer covariance pairs were calculated at the longer separation distances, increasing the noise in the calculation. The curves show a correlation which drops off roughly

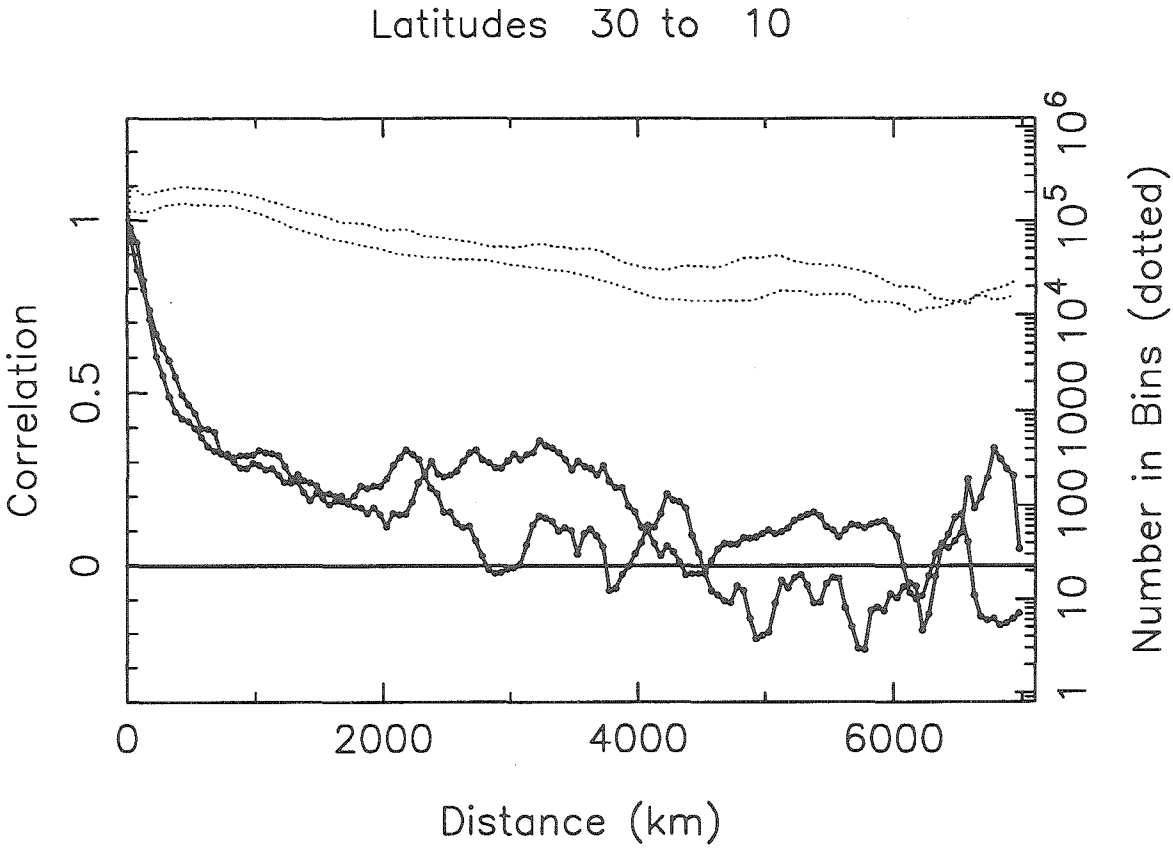


Fig. 3.5: Correlation as a function of separation distance for the latitude bin centered on 20°N. The solid lines are the correlations computed for the two groups of data around $L_S \sim 0°$. The dotted lines represent the log of the number of individual covariances averaged in each 50 km bin for the two groups. There are roughly 10^5 individual covariances in each bin, reducing the noise on each correlation curve. The correlation drops like an exponential with a characteristic length of order 1000 km.

exponentially, with a characteristic length scale of order 1000 km, similar to the simple estimates presented above.

Figure 3.6 shows these same correlation curves for all of the latitude bins for the first and second groups of correlations around $L_S \sim 0°$. Again, the two solid lines are not identical but follow the same general trends. In general the correlations are all falling off roughly exponentially with a characteristic length scale of order 1000 km. However, there do appear to be significant differences in the curves between different latitudes. These differences are in the shape of the correlation and also in its overall length scale. We do not attempt to quantify or explain the shape differences here, as those are at least in part due to noise in the correlation calculations.

We attempted to fit exponentials of the form $\rho = e^{-kd}$, where k is a fit parameter and d is again the separation distance. We weighted the fits by the number of covariance pairs in each bin to produce a weighted least squares fit to each of the curves. For the $L_S \sim 0°$ subset, this yielded values for each 20° latitude bin from north to south of 240km, 700km, 1300km, 1200km, 1700km, 1500km, 2600km, 1500km, and 2400km respectively. The numbers listed here are an average of the two results from the first and second groups analyzed at $L_S \sim 0°$ except in cases where the correlation curves were poorly fit by an exponential. An example of this is the upper curve for the latitude bin centered on 40°N where the correlation curve is essentially constant out to 4000km.

We also tried to quantify the correlation length scales by integrating the area under the correlation curves out to their first zero crossing. The results of this attempt were less useful than the exponential fits, as the noise on the

Fig. 3.6: Correlation as a function of separation for all of the latitude bins for the two groups of data analyzed around $L_S \sim 0°$. The structure of each plot is the same as that of Fig. 3.5. Note the similarities between the two solid lines in each plot, and the differences in overall length scale and shape of the correlation curves among the different plots.

Latitudes 90 to 70
Latitudes 70 to 50
Latitudes 50 to 30
Latitudes 30 to 10
Latitudes 10 to −10
Latitudes −10 to −30
Latitudes −30 to −50
Latitudes −50 to −70
Latitudes −70 to −90
Correlation
Number in Bins (dotted)
Distance (km)
0
0.5
1
2000
4000
6000
10
100
1000
10^4
10^5
10^6

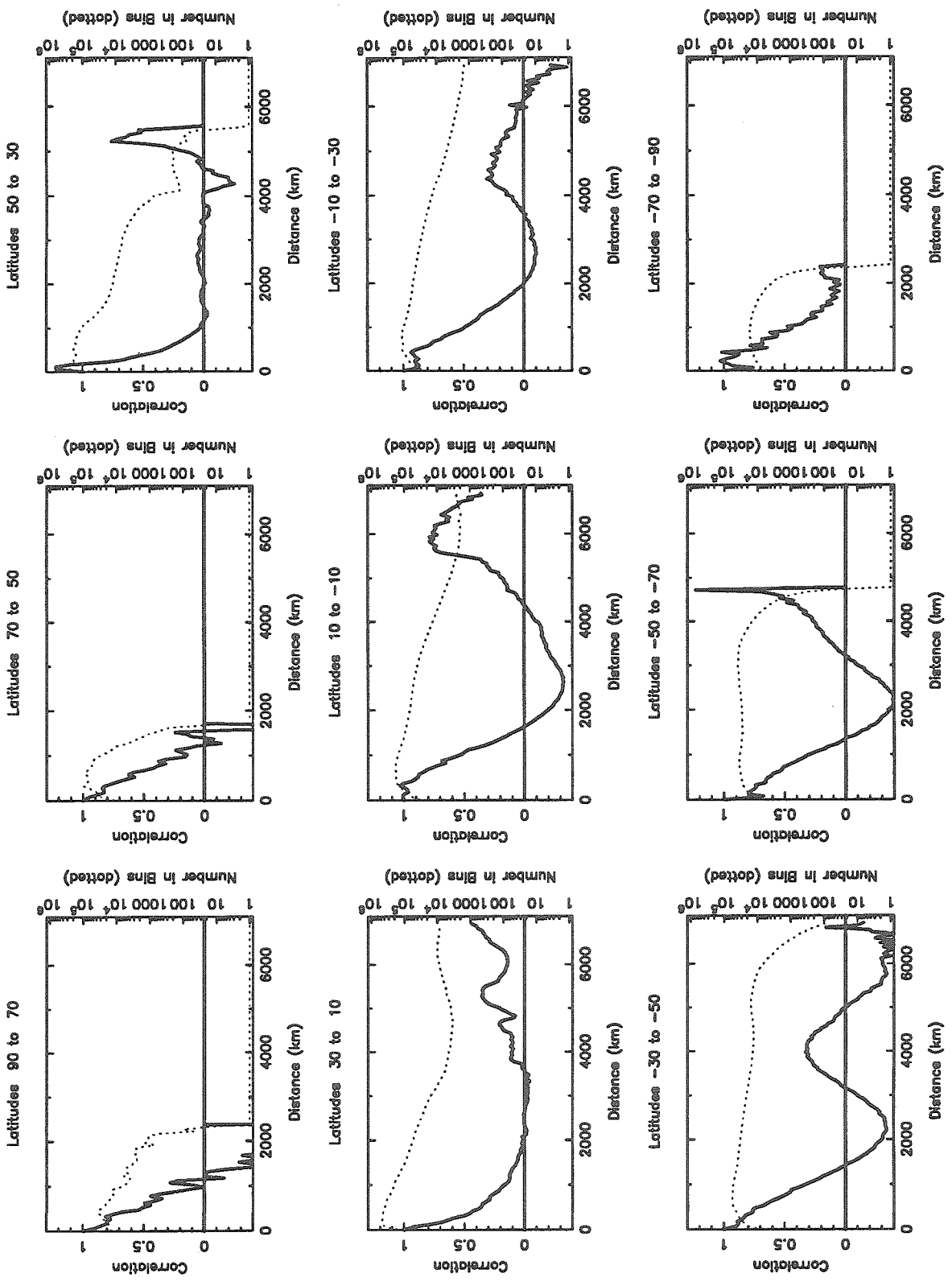
correlation curves significantly influenced the overall result. We present no results from this attempt.

Perhaps the most infallible method is simply to estimate 'by eye' the exponential length scale one would fit to the correlation curves. In doing so, we note similar trends as found with the least squares fits. Going from north to south, values of about 250km, 600km, 1500km, 700km, 1300km, 1400km, 2500km, 1700km, and 1800km are found. Note that the northernmost correlation curves seem to have a significantly shorter length scale than the others during this season.

Figure 3.7 shows the correlation curves for the northern fall equinox season ($L_S \sim 180°$). In this case, we only computed one group of correlations for this season, so the errors in these correlations can only be estimated from those found in the opposite season. As the coverage was somewhat worse during the northern fall equinox, the errors are probably slightly larger for this subset. Nevertheless, one again sees similar general behavior in these plots, namely, a drop-off in roughly 1000 km, perhaps being shorter in the north. However, there are features present at this season not observed in the opposite season. From the equator southwards (panels 5, 6, 7, and 8), the correlation curves are not simply exponential in form. They seem to indicate a dominant mode of oscillation, in addition to the general decrease in correlation with separation distance. At the equator, this oscillation has a wavelength of roughly $2\pi \times 1000$ km. At 20°S, the wavelength is $\sim 2\pi \times 700$ km. Between 40°S and 60°S the wavelength is roughly $2\pi \times 640$ km. These correspond to zonal wavenumbers of 3-4 at these latitudes.

With the oscillations present in the southern correlation curves for $L_S \sim$

Fig. 3.7: Correlation as a function of separation for all of the latitude bins for the data analyzed around $L_S \sim 180°$. The structure of each plot is the same as that of Fig. 3.5. Note the general fall-off in about 1000km, and the wave-like structure suggested in the southern plots. Compare this figure with Fig. 3.6.


Latitudes 90 to 70
Latitudes 70 to 50
Latitudes 50 to 30
Latitudes 30 to 10
Latitudes 10 to −10
Latitudes −10 to −30
Latitudes −30 to −50
Latitudes −50 to −70
Latitudes −70 to −90
Correlation
Number in Bins (dotted)
Distance (km)
0
0.5
1
10
100
1000
10^4
10^5
10^6
2000
4000
6000

180°, we did not attempt to fit those curves to simple exponentials, $\rho = e^{-kd}$. However, for the northern correlation curves for $L_S \sim 180°$ we found characteristic length scales of 600km, 750km, 400km, and 550km for the latitude bins centered at 80°N, 60°N, 40°N and 20°N respectively. Using the 'by eye' fits, these same general length scales are found for the north during this season. One can also estimate 'by eye' probable characteristic correlation length scales for the southern curves in this season, although with quite large errors. We estimate 2500km for the equator, 1800km for 20°S, 3000km for 40°S, 3000km for 60° and about 1100km for 80°S. These estimates again support a shortening of the length scale towards to north as was seen during the $L_S \sim 0°$ season.

We have compiled these weather correlation length scale determinations in Fig. 3.8. For the $L_S \sim 0°$ results (solid lines), which show a significant range of values for each latitude bin, we have indicated the rough extremes as found through the two techniques (i.e., exponential fits and 'eye' fits) for the two subsets analyzed. For the $L_S \sim 180°$ results (dotted line) we only show a single point for each latitude bin as the two estimation techniques gave similar values for the one subset analyzed in this season. In this plot, it is clear that the northern latitudes consistently show smaller weather correlation length scales, while the south exhibits generally larger length scales.

We also examined the magnitude of the weather temperature variance. Figure 3.9 depicts the weather temperature variance as a function of latitude for the two groups analyzed around $L_S \sim 0°$ (solid line) and the one group centered on $L_S \sim 180°$ (dotted line). The differences between the solid curves are indicative of the errors present in this calculation, which are significant.

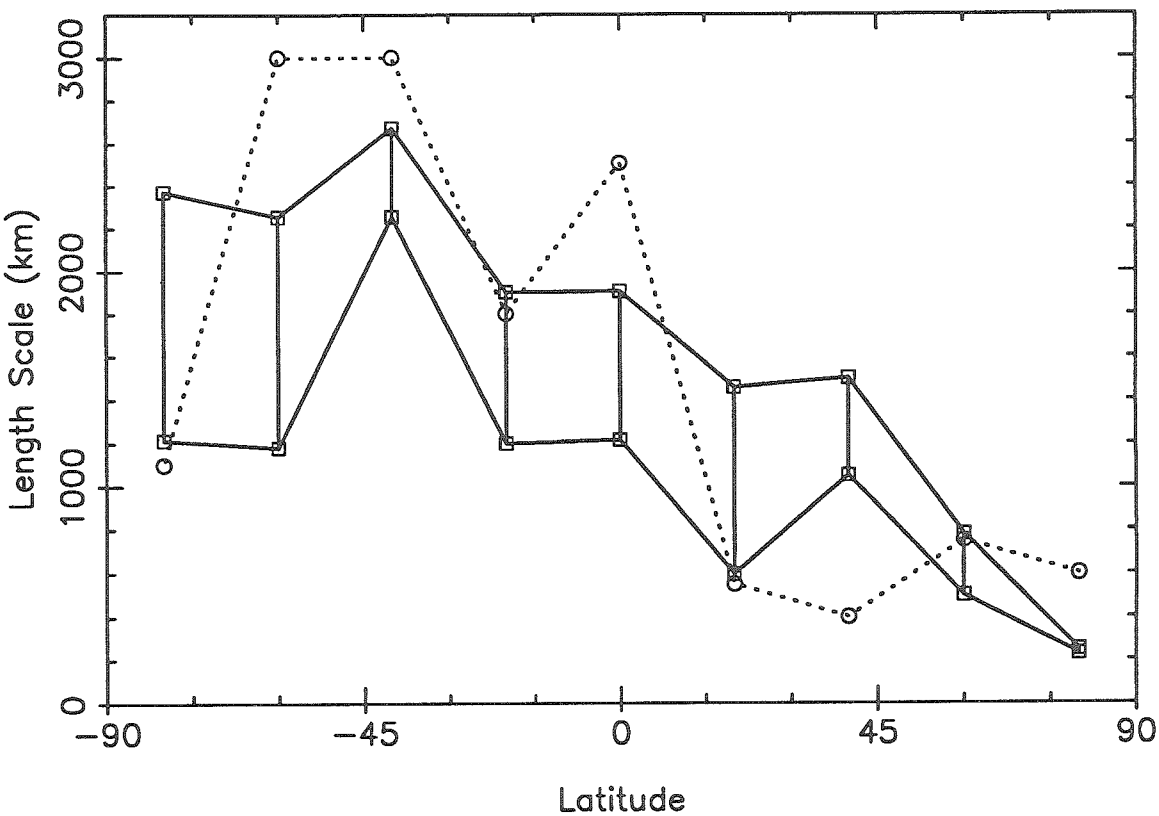


Fig. 3.8: Weather temperature correlation length scales as a function of latitude for $L_S \sim 0°$ (solid lines) and $L_S \sim 180°$ (dotted line). The vertical solid lines connect the extrema of the length scales found for each latitude bin for the $L_S \sim 0°$ subsets, showing roughly the precision of these values. Only one subset of $L_S \sim 180°$ was analyzed so no ranges are shown for those values. All these values were determined by fitting exponentials (mathematically or by 'eye') to the correlation curves in Figs. 3.6 and 3.7. Note the smaller length scales in the north.

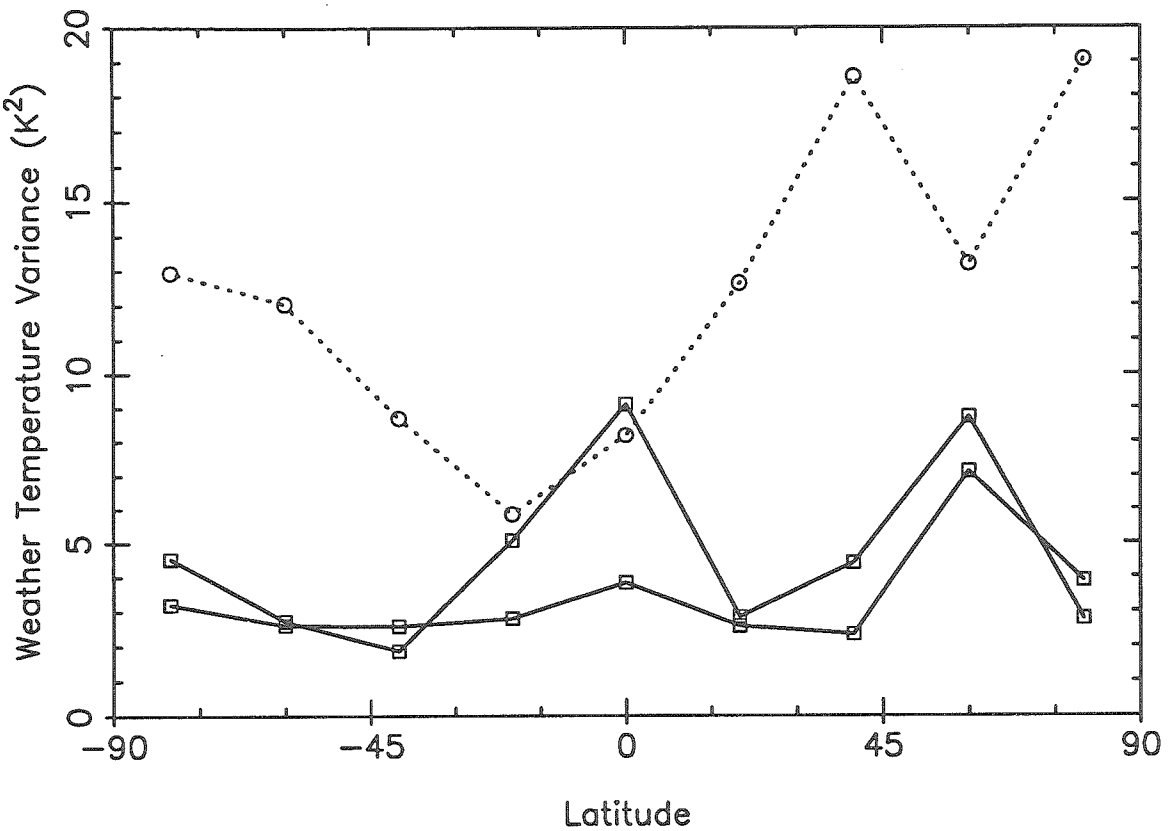


Fig. 3.9: Weather temperature variance as a function of latitude for the two groups around $L_S \sim 0°$ (solid lines) and the group around $L_S \sim 180°$ (dotted line). These curves represent the variance in the temperature at the 0.6mbar level caused by weather phenomena on Mars. They are computed by requiring the extrapolated weather temperature correlations at zero separation to be 1.0, then adjusting the variance to achieve this. The differences between the two solid curves are indicative of the errors in these results, which are significant. Thus, only global averages should be considered. For the two groups at $L_S \sim 0°$, values of $3 - 5K^2$ are found, while for $L_S \sim 180°$, a value of about $11K^2$ is found.

Table 3.I
Martian Weather Correlation Length Scale Estimates

Technique	Length (km)
Theory	$L_D(45°) \sim 1000$
Lander $V \cdot \Delta t$	~ 3000
Lander Δk	1000-2000
This Work	250-3000

The latitudinal structure in these curves is mostly due to noise, so only global averages should be considered. Thus, a weather temperature variance of between $3 - 5K^2$ is found as an area-weighted global average for $L_S \sim 0°$. At $L_S \sim 180°$, a value of $\sim 11K^2$ is found.

3.5 Discussion

In this section, we summarize our determination of the weather temperature correlation length scale. We also suggest some explanations for details observed in the correlation curves, and what that might suggest about the weather of Mars. Furthermore, we use our length scale estimates to suggest a number of landers necessary to observe well the global atmospheric state of Mars. Finally, we compare our determinations of the weather temperature variances with published results from GCM simulations and other data sets.

We have summarized our weather correlation length scale determination from IRTM data and estimates from other techniques in Table 3.I. The results are quite consistent, suggesting a length scale on the order of 1500km globally,

with hints of seasonal and latitudinal differences. There are indicators in the Viking Lander data estimates that the length scale may be different in the zonal direction than it is in the meridional direction, although our IRTM analysis could not be used to address this question. These results can be compared with the published results from GCM calculations reported in Barnes *et al.* (1993). They provide wavenumber-frequency spectra for the transient eddies found in the Ames GCM simulations for several seasons and at particular latitudes. Their spectra have similar widths in wavenumber and frequency as the Viking results presented in Barnes (1980,1981). It was these spectra from Barnes (1980,1981) which we used to estimate the weather correlation length scale from Viking Lander data, and which we found to be consistent with our IRTM analysis. Therefore, we also find good accord between our IRTM results and those of the Ames GCM simulations.

Why all of these different data sets should yield the same weather correlation length scales is an interesting question of itself. The Landers sampled the surface pressure, while the IRTM T15 was measuring temperature at about an altitude of 25km. The fact that they yield similar results is indicative that the weather phenomena are similarly manifested in both pressure and temperature variations for Mars. It also implies that the weather phenomena investigated have large vertical extents, i.e., extending over at least two scale heights. This hypothesis is supported by GCM calculations reported in Barnes *et al.* (1993). In that work, they also discuss the vertical structure of the eddies observed in their model runs. They find eddies which are vertically coherent over the full domain of the model, namely about 4 scale heights. This behavior was also previously observed in the Martian

atmosphere by Conrath (1981) who analyzed Mariner 9 IRIS spectra. He observed an atmospheric disturbance which had both a consistent frequency spectra and phase structure with height up to about 4 scale heights as well. Therefore, we do not find it surprising that the length scales we derive from these different data sets (and regions of the atmosphere) are all consistent.

We noted an apparent latitudinal dependence in our estimates of the weather temperature correlation length scales (see Fig. 3.8). This was manifested as a quickly decreasing length scale toward the north pole. Such a phenomenon was noted in the Earth's atmosphere by Ghil *et al.* (1979) who examined prediction error correlations for use in data assimilation. They explained it in terms of the variation of the Rossby radius with latitude. Recall that the Rossby radius can be expressed by $L_D = NH/2\Omega \sin\phi$, and so decreases with increasing latitude (ϕ). This phenomenon has also been observed in eddies in the Earth's oceans as reported in Krauss *et al.* (1990). Thus, we do not find it surprising to see a variation with latitude of the weather temperature correlation length scale that roughly echoes the variation of the Rossby radius.

It is interesting, however, that we do not see a symmetrical distribution of length scales about the equator in these equinoctial observations. Although the southern-most latitude bin also showed slightly shortened length scales, the decrease was much less dramatic than in the north. This trend was the same for both $L_S \sim 0°$ and $L_S \sim 180°$, not reversing between fall and spring as might be expected. The retreating polar cap during the equinoxes switches hemispheres between these two time periods, namely being in the north for $L_S \sim 0°$ and in the south for $L_S \sim 180°$. Thus, the hemispheric asymmetry

of the weather correlation length scales is not just a seasonally controlled phenomena, and is perhaps due to persistent hemispheric differences in the weather. Examples of potential persistent influences on the weather correlation length scale could be topography or thermal inertia differences between the hemispheres. These could influence the regions in which weather systems are generated, and thus their length scales. For example, if the northern latitudes during $L_S \sim 0°$ and $L_S \sim 180°$ are dominated by weather generated locally (i.e., waves centered on far-north latitudes), then we would expect them to exhibit the shorter length scales observed. Similarly, if the southern latitudes during these seasons are dominated by larger, nearly planet-scale weather systems generated nearer the equator, we would expect them to exhibit longer, more consistent length scales throughout the southern hemisphere. Correlations computed in that region would sample the same, large weather system. Note, however, that this is simply one of many possible self-consistent explanations of the hemispheric differences we observe in the length scales during both seasons.

During $L_S \sim 180°$, we also noted a dominant oscillation in the atmosphere in the southern hemisphere of zonal mode 3-4. This is similar to the observations of Barnes (1980,1981), and the GCM results reported in Barnes *et al.* (1993), although they found dominant wavenumbers of between 2 and 4 during this season. Our analysis seems to quite strongly support the higher wavenumbers reported by Barnes *et al.* (1993) from the GCM. Thus, most of the gross structures found in our correlation analysis can be explained within the current understanding of the nature of the transient eddies present in the Martian atmosphere.

Seaman (1977) showed that the accuracy in estimating the atmospheric state from a network of landers is directly related to the spacing of landers divided by the weather correlation length scale. Thus, one can estimate the number of landers needed on Mars to observe the global surface pressure without missing any significant features in the weather signal. We will take the smallest length scale measured in each latitude bin for either season investigated and assume that that is the separation of landers in that latitudinal bin. Seaman's (1977) results show that this is a good choice for lander spacing. Performing this yields a total number of landers of roughly 110, yet with only 10 south of the equator. However the northern mid-latitudes require some 70 landers because of the extremely small (i.e., 400km) length scales observed there in this analysis. Calculating the number of landers needed from the smallest length scales found might seem like a worst-case scenario, but if complete resolution is desired, then this number is required. Furthermore, this number could grow if even smaller length scales were found upon analyzing the other Martian seasons.

The weather temperature variances computed in our IRTM analysis can be also compared with other published work. Barnes *et al.* (1993) report the GCM's transient eddy temperature variance as a function of latitude and altitude. At $L_S \sim 0°$ (their Fig. 18), near 25 km altitude, they report variances of between $0.5K^2$ and $3K^2$, with a globally averaged value of about $1K^2$. Our results were consistently higher than this, with a globally averaged value of about $4K^2$ for this season. Nevertheless, the match is good on an order of magnitude level. At $L_S \sim 180°$ (their Fig. 16), near 25 km altitude, they report variances of between $0.5K^2$ and $11K^2$, although values

above $4K^2$ are found only between 40°N and 65°N. A globally averaged value of their findings is roughly $2.5K^2$. For this season, we again found values consistently higher than this, with a globally averaged value of about $11K^2$. While we caution against interpreting the latitudinal structure in our weather temperature variance results, for $L_S \sim 180°$, we do see somewhat greater values north of 40°N. This hints at an agreement with the GCM results, in spite of the magnitude differences.

As another comparison for our variance calculations, Conrath (1981) also computed a temperature amplitude from a wave he observed in Mariner 9 IRIS data. While his data set is from a time slightly earlier ($L_S \sim 340°$) than our northern spring equinox subset ($L_S \sim 0°$), it is still useful to compare against. He found wave amplitudes of up to 4K at the 25 km altitude, in a latitudinally confined region between about 50°N and 70°N. This observation is more consistent with the GCM results than our observations, however the agreement with our analysis is not particularly bad either. One possible explanation for the discrepancy between our analysis and the Conrath and GCM results could be interannual variability. This could be addressed with GCM simulations, although to date, no multi-year statistics have been reported in the literature. Nevertheless, we at least see an agreement in the order of magnitude of the weather temperature variances revealed in these three analyses.

3.6 Summary

In analyzing the Viking IRTM T15 data sets to derive the weather temperature correlation length scales, we have established another benchmark with which to compare the current crop of Martian GCM's to the realities of the Martian atmosphere. We found consistent length scales for the correlation of roughly 1500 km, and hints that it may have some interesting latitudinal structure (shorter in the north). We found weather temperature variances at 25 km altitude of order $4K^2$ ($L_S \sim 0°$) and $11K^2$ ($L_S \sim 180°$), somewhat higher than that reported by others (e.g., Barnes *et al.* 1993, Conrath 1981). The in-depth analysis on the IRTM data was only performed during the non-dusty times around spring and fall equinox. However, both Lander observations and GCM calculations suggest that perhaps the most interesting time to study weather phenomena is during the winter season, when the temperature variances increase by perhaps a factor of two. Therefore, to add to this work, to better constrain the number of meteorological landers needed at Mars and to further understand the nature of weather on Mars, we hope to eventually perform this type of study throughout the whole of the IRTM data set. Furthermore, the weather correlation length scales derived here should be useful in developing the first operational atmospheric data assimilation systems for Mars, when global data becomes available. Finally, these length scales should be used as a guide in determining the density of landers required to compile a useful meteorological network for Mars.

3.7 References

BANFIELD, D., A.P. INGERSOLL, AND C.L. KEPPENNE 1994. A steady-state Kalman filter for assimilating data from a single polar-orbiting satellite. submitted to *J. Atmos. Sci.*.

BARNES, J.R. 1980. Time spectral analysis of midlatitude disturbances in the Martian atmosphere. *J. Atmos. Sci.* **37**, 2002- 2015.

BARNES, J.R. 1981. Midlatitude disturbances in the Martian atmosphere: a second Mars year. *J. Atmos. Sci.* **38**, 225- 234.

BARNES, J.R. 1984. Linear baroclinic instability in the Martian atmosphere. *J. Atmos. Sci.* **41**, 1536-1550.

BARNES, J.R., J.B. POLLACK, R.M. HABERLE, C.B. LEOVY, R.W. ZUREK, H. LEE, AND J. SCHAEFFER 1993. Mars atmospheric dynamics as simulated by the NASA Ames general circulation model 2. Transient baroclinic eddies. *J. Geophys. Res.* **98**, 3125-3148.

CHASE, S.C., JR., J.L. ENGEL, H.W. EYERLY, H.H. KIEFFER, F.D. PALLUCONI, AND D. SCHOFIELD 1978. Viking infrared thermal mapper. *Appl. Opt.* **17**, 1243-1251.

CONRATH, B.J. 1981. Planetary-scale wave structure in the Martian atmosphere *Icarus* **48**, 246-255.

GHIL, M., M. HALEM, AND R. ATLAS 1979. Time-continuous assimilation of remote-sounding data and its effect on weather forecasting. *Mon.*

Wea. Rev. **107**, 140-171.

HABERLE, R.M., J.B. POLLACK, J.R. BARNES, R.W. ZUREK, C.B. LEOVY, J.R. MURPHY, H. LEE, AND J. SCHAEFFER 1993. Mars atmospheric dynamics as simulated by the NASA Ames general circulation model. 1. The zonal mean circulation. *J. Geophys. Res.* **98**, 3093-3123.

HAMILTON, K., AND R.R. GARCIA 1986. Theory and observations of the short-period normal mode oscillations of the atmosphere. *J. Geophys. Res.* **91**, 11867-11875.

KAHN, R. 1984. The spatial and seasonal distribution of Martian clouds and some meteorological implications. *J. Geophys. Res.* **89**, 6671-6688.

KIEFFER, H.H., P.R. CHRISTENSEN, T.Z. MARTIN, E.D. MINER, AND F.D. PALLUCONI 1976. Temperatures of the Martian surface and atmosphere: Viking observations of diurnal and geometric variations. *Science* **194**, 1346-1351.

KRAUSS, W., R. DÖSCHER, A. LEHMANN, AND T. VIEHOFF 1990. On eddy scales in the eastern and northern Atlantic ocean as a function of latitude. *J. Geophys. Res.* **95**, 18049-18056.

MARTIN, L.J., AND R.W. ZUREK 1993. An analysis of the history of dust activity on Mars. *J. Geophys. Res.* **98**, 3221-3246.

MARTIN, T.Z., AND H.H. KIEFFER 1979. Thermal infrared properties of the Martian atmosphere 2. The 15-μm band measurements. *J. Geophys. Res.* **84**, 2843-2852.

SANTEE, M.L., AND D. CRISP 1992. Thermal structure and dust loading of the Martian atmosphere during late southern summer: Mariner 9 revisited. *J. Geophys. Res.* **98**, 3261-3279.

SEAMAN, R.S. 1977. Absolute and differential accuracy of analyses achievable with specified observational network characteristics. *Mon. Wea. Rev.* **105**, 1211-1222.

TILLMAN, J.E., R.M. HENRY, AND S.L. HESS 1979. Frontal systems during passage of the Martian north polar hood over the Viking Lander 2 site prior to the first 1977 dust storm. *J. Geophys. Res.* **84**, 2947-2955.