Evaluation of tropospheric water vapor profiling using eye-safe, infrared differential absorption lidar

Barry J. Rye,¹ Janet L. Machol,² Christian J. Grund,² R. Michael Hardesty²

¹ Cooperative Institute for Research in Environmental Sciences
University of Colorado / NOAA, Environmental Technology Laboratory
325 Broadway, Boulder, CO 80303

² NOAA / ERL / Environmental Technology Laboratory
325 Broadway, Boulder, CO 80303

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Report for Department of Energy Atmospheric Radiation Measurement Program, supported under interagency agreement DE-AI-03-94-ER61761

May 14, 1996

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED
Evaluation of tropospheric water vapor profiling using eye-safe, infrared differential absorption lidar

Summary

Continuous, high quality profiles of water vapor, free of systematic bias, and of moderate temporal and spatial resolution are fundamental to the success of the ARM CART program. In addition, these should be acquired over long periods at low operational and maintenance cost. The development and verification of realistic climate model parameterizations for clouds and net radiation balance, and the correction of other CART site sensor observations for interferences due to the presence of water vapor are critically dependent on water vapor profile measurements. To date, application of profiles have been limited by vertical resolution and uniqueness of solution (e.g. high resolution infrared (IR) Fourier transform radiometry), poor spatial and temporal coverage and high operating cost (e.g. radiosondes), or diminished daytime performance, lack of eye-safety, and high maintenance cost (e.g. Raman lidar). Recent developments in infrared laser and detector technology make possible compact IR differential absorption lidar (DIAL) systems at eye-safe wavelengths. In the studies reported here, we develop DIAL system performance models and examine the potential of solving some of the shortcomings of previous methods using parameters representative of current technologies. These simulations are also applied to determine the strengths and weaknesses unique to the DIAL method for this application.

Two requirements for routine monitoring of tropospheric water vapor are (i) a transceiver with enough mean power to make unbiased and sufficiently high precision measurements of the returns within the atmospheric fluctuation times and (ii) eye-safety. Because eye safety constraints are less severe in the IR, especially beyond ~1.5 μm, the initial study, Part I, concentrates on examining the potential infrared lasers operating at 2 and 10 μm which are currently available in our laboratory. Because of detector limitations in the IR, these systems require heterodyne detection, a technique which has seldom been used with DIAL. The intrinsic advantages of heterodyning are a small field of view and bandwidth which substantially reduce the background, a noise source that limits Raman lidar largely to nighttime operation. Heterodyning requires the use of a low peak power, high pulse repetition frequency (prf) laser systems to average speckle. The average powers of the 2 and 10 μm transmitters are 5 and 10 W, respectively, but adequate data could be obtained from a system with much lower power.

In Part II, we consider systems at shorter wavelengths where direct detection is applicable. In these systems, the background signal can still be reduced to acceptable levels by limiting the field of view. Simulations are used to examine the effects of various parameters such as pulse power, wavelength, prf, and method of detection on the data accuracy. A comparison is made of three detection techniques: heterodyning and direct detection with either pulses or pseudo-random code modulation.

Recent advances in solid-state laser and detector technologies are considered which may permit the development of efficient and compact lidar systems at wavelengths between 0.8 and 1.5 μm. In particular, diode lasers are considered as simple sources of tunable, single-frequency light, and several laser designs are suggested for the water vapor DIAL system. The stability, tunability, and spectral characteristics of diode lasers with respect to DIAL water vapor measurements are assessed in laboratory experiments. Other required optical elements are also examined.

It is concluded that all of the considered systems would provide adequate resolution for water vapor DIAL measurements in the lower troposphere. The heterodyne technique is better at longer wavelengths and direct detection is preferred at shorter wavelengths. The crossover point is at ~2μm. The available lidars discussed in Part I will be converted to DIAL systems and tested for their capabilities in profiling water vapor. A pulsed direct detection system using diode-based lasers is the most promising choice for a simple, compact and relatively low cost ground-based DIAL system, but will require about one year’s worth of development.
Table of Contents

Summary

Part I: Available lidar technology

1 Motivation
   1.1 The importance of water vapor to climate assessment 1
   1.2 Water vapor measurements at CART 1
   1.3 Differential absorption lidar (DIAL) 3
   1.4 Layout of this report 4

2 General properties of the DIAL technique 5

3 Vertical profiling of water vapors
   3.1 Comparison with horizontal and near-horizontal profiling 8
   3.2 CO₂ heterodyne lidar simulations 9
   3.3 2 μm lidar simulations - heterodyne and direct detection 10
   3.4 Temperature dependence 10

4 Turn-key technology assessment
   4.1 Introduction 11
   4.2 2 μm wavelength region
      4.2.1 Background 12
      4.2.2 Wavelength 12
      4.2.3 Alignment stability 13
      4.2.4 Detector linearity 14
      4.2.5 Maintenance and operational requirements 14
      4.2.6 Calibration requirements 15
   4.3 10 μm wavelength region
      4.3.1 Background 15
      4.3.2 Wavelength stability 16
      4.3.3 Alignment stability 16
      4.3.4 Detector linearity 16
      4.3.5 Maintenance and operational requirements 17
      4.3.6 Calibration requirements 17
   4.4 Data acquisition processor technology 18

5 Appendix 1: DIAL equations
   5.1 Precision of DIAL estimates - the lidar equation 18
   5.2 Direct detection and heterodyne lidar 20
   5.3 Heterodyne lidar antennas
      5.3.1 Introduction 23
      5.3.2 Gaussian beam approximation 25
      5.3.3 Refractive turbulence 26
5.3.4 Summary
5.4 CNR, SNR, and fractional standard deviation
5.5 Return power estimation in heterodyning
5.6 Column content and range-resolved concentrations

6 Appendix 2: Details of the simulations
   6.1 Geometry
   6.2 Backscatter coefficient
   6.3 Refractive turbulence
   6.4 Physical parameters and species concentrations
   6.5 Line absorption
   6.6 Water vapor continuum absorption at 10 μm
   6.7 Examples of absorption spectra
   6.8 Summary of lidar parameters

7 Conclusions

References for Part I

Part II: Diode-based lidars

8 Introduction

9 Simulation
   9.1 Parameters
   9.2 Bias and atmospheric variability
   9.3 Simulation results
   9.4 Comparison of rm-code and pulsed direct-detection lidars
   9.5 Comparison of heterodyne and direct-detection lidars
   9.5 Conclusions

10 System design
   10.1 Detectors
   10.2 Lasers
      10.2.1 All-semiconductor laser
      10.2.2 Diode laser with fiber amplifier
      10.2.3 Other solid state laser options
   10.3 Other system components

11 Laboratory Studies

12 Conclusion

References for Part II
Part I: Evaluation of tropospheric water vapor profiling using eye-safe, infrared differential absorption lidar:

Existing lidars

1 Motivation

1.1 The Importance of Water Vapor to Climate Assessment

Modulations of the radiative divergence of the cloud-free atmosphere are largely controlled by the vertical distribution of water vapor. As a greenhouse gas, water vapor directly reduces the outgoing infrared (IR) flux to a degree dependent on both temperature (correlated with altitude) and concentration. Good vertical resolution and concentration accuracy are vital to the assessment of this effect. Indirectly, hygroscopic aerosols swell with increasing relative humidity, often causing marked short wavelength local albedo changes, reducing the effective solar heating. Clouds have been identified as potent climate modulators. The vertical and regional distribution of cloudiness, and cloud morphology (an important factor in cloud impacts on radiative transfer, Stephens and Greenwald, 1991) are also largely dependent on the availability of water vapor.

In addition to the direct radiative effects, there are many indirect mechanisms whereby the distribution and transport of water vapor may have large impacts on climate and on radiative balance. For example, water vapor provides a medium by which vast quantities of heat may be absorbed and stored without sensible temperature change, only to be transported regionally or globally, then released, modifying large scale circulation patterns and local radiative balance. Water vapor plays a key role in the transport and storage of heat in the oceans through precipitation and consequent formation of stable salinity gradients, ultimately modifying surface temperature distributions. Water vapor transport into otherwise dry polar regions modifies radiative balance through the deposition of high albedo, high emissivity insulating layers of snow, which subsequently return to the ocean as glacial fresh water outflows, further modifying ocean currents and surface temperature distributions. Observing and understanding the factors controlling water vapor flux modulations between the atmosphere, land, and ocean surfaces is critical to the assessment of these indirect mechanisms on radiation and climate.

The development of accurate model parameterizations for the direct and indirect impacts on climate of the global redistribution of water vapor resulting from possible climate change scenarios depends upon a complete understanding of the atmospheric and biogeochemical factors controlling the distribution of the species, on environmental interactions that can radically alter the impact of water vapor on the radiative balance (e.g. swelling aerosols, cloud formation and morphology), and on understanding the water vapor flux and transport mechanisms in the global water cycle. Accurate vertically resolved water vapor profile measurements throughout the troposphere are essential to continuing development of the understanding of these factors, and the resultant improved model parameterizations.
1.2 Water Vapor Measurements at CART

An ambitious central objective of the CART site strategy is to approach complete characterization of the radiative divergence and content of a well defined atmospheric column through the use of complimentary remote-sensing and in-situ instrumentation. Time series of these measurements in different climatic through seasonal and diurnal instrumentation are intended to provide inputs for the development and verification of climate model parameterization, and to provide a consistent database from which advanced climate research may proceed. As recognized by the ARM program, of key importance to this suite of instrumentation is the continuous measurement of high accuracy water vapor profiles with temporal and spatial resolution appropriate to study significant radiative modulations and structural scales. To be economically practical, the instrumentation should eventually be capable of routine operation with minimal attention by technical personnel of moderate skill. What is needed to fulfill CART site goals is a means of acquiring continuous, high quality vertical profiles of water vapor, free of bias and systematic error, and of moderate temporal and spatial resolution, over long periods with low operational and maintenance cost.

In-situ sensors on towers can provide the needed temporal and vertical resolution, but cannot reach above the lowest regions of the typical daytime boundary layer. Tethered balloons can reach through a larger portion of the boundary layer, but are limited by atmospheric conditions, pose air traffic hazards, and are costly to outfit and maintain with sufficient vertical resolution. Radiosondes become expensive if launched at adequate sampling intervals, and sample only the air along the irregular wind-drifted path through which they are ascending. They have limited accuracy at low temperatures, and exhibit significant hysteresis. Operation of airborne sensors is quite expensive, and can thus be considered only for brief intensive measurement periods.

Remote sensors can potentially provide the needed vertical and temporal resolution on a continuous basis. Up-looking ground-based passive radiometers such as Fourier transform infrared (FTIR) spectrometers, have recently been shown to be capable of reasonable clear air resolution in the boundary layer under ideal conditions, but have poor vertical resolution above 2 km. These systems have difficulty whenever thin clouds or turbid aerosol conditions perturb radiances within the instrument field of view, and may require multi-species retrievals to eliminate interferences in polluted atmospheres. Raman lidars have been developed which provide the needed accuracy and vertical resolution. Although intrinsically simple, Raman lidars have proven larger and more cumbersome to operate than anticipated, they usually depend upon excimer lasers requiring significant maintenance and which present potential hazards in continuous operation, and have only proven feasible for profiling to a few km altitude in the daytime. These problems are largely due to the small Raman scattering cross sections and the large signal dynamic range, caused partly by the large ozone attenuation and Rayleigh scattering cross section at the ultraviolet (UV) wavelengths employed.

As an alternative, differential absorption lidar (DIAL) techniques operating in the infrared can also be employed to achieve the needed range resolution with better temporal resolution than practical Raman systems, while remaining immune to daytime skylight noise. A comparative review of DIAL and Raman lidar has recently been published by Grant (1991). DIAL techniques provide better range resolution than Fourier transform infrared radiometric (FTIR) profilers without the sensitivity to
aerosols and clouds or problems related to uniqueness of solution. With ARM CART goals in mind, this study examines and compares the theoretical feasibility of performing water vapor DIAL measurements at 2 μm and 10 μm wavelengths and provides realistic error estimates for practical systems. These wavelengths were chosen because laser technology suitable for near turn-key operation is at hand, and because these wavelengths are unconditionally eye-safe for the transmitted pulse energies considered. As a unique aspect of this study, we examine and compare the sensitivity of coherent and direct detection techniques for DIAL measurements, the impact of detection method on the laser requirements, and the maturity of technologies required to achieve near turn-key operation suitable for CART.

1.3 Differential Absorption Lidar (DIAL)

The DIAL technique has been implemented to determine the concentration of water vapor and many other species at many laboratories both in the United States and abroad since the pioneering demonstrations of Schottland (1966) which commenced using searchlights a few years before the laser was invented. In its simplest form, DIAL is based on comparing the transmission through the atmosphere of beams at two wavelengths which are absorbed differently by the species being monitored. The ratio of the attenuations of the two beams is given by the Bouguer-Lambert-Beers law as $\exp[-n \sigma r]$ where $n$ and $\sigma$ are the concentration and differential absorption cross-section of the specie and $r$ is the path length. In a lidar, the transmitted beams are scattered back to the receiver by the atmosphere, resulting in a measurement of the column content $2nr$ within the identical outgoing and return paths along the line of sight; because backscatter is distributed along the transmitted beam path, returns over different path lengths $r$ can be compared to determine the range resolved concentration. The precision with which the attenuation at the two wavelengths can be compared is determined by the signal-to-noise ratio (SNR) - the ratio of the return signal power in the lidar to the power of the noise in the receiver output - which is therefore the principal statistical figure of merit for a DIAL system.

Most DIAL systems used for monitoring water vapor have operated at a wavelength of about 700 nm (Browell et al., 1979, Cahen et al., 1982, Zuev et al., 1983, Bosenberg, 1983, but recent developments in infrared technology make possible compact systems at eye-safe wavelengths. The problems with earlier infrared lidars which used direct detection were in part that photodetectors in the infrared are relatively insensitive. The principal noise contributions in the SNR arose from background radiation on the photodetector and thermal noise, and large, powerful, lasers were needed to act as transmitters. These difficulties can be overcome by use of heterodyne receivers. Heterodyne receivers are insensitive to thermal background light and are essentially photon-limited so that transmitted pulses of rather low energy can provide adequate returns. Their principal noise source is fluctuation in the return signal itself, which is known as coherent fading or speckle. A heterodyne lidar must be designed to average this out. Operation at a high pulse repetition frequency (PRF) can provide the necessary averaging, but only recently have suitable laser sources become available.

There are two problems fundamental to DIAL that especially relate to the problem of vertical profiling from the ground. The first is that the technique benefits from both strong returns and strong
absorption. If the absorber is water vapor and the lidar line of sight points upwards from the ground, both the absorption and the backscatter will reduce rapidly with altitude. At high altitudes the absorption could be increased by use of a strong molecular absorption line, but this leads to a weaker return due to stronger low altitude attenuation and a lower SNR. To some extent this can be countered in a heterodyne system because the return can be increased from any particular range by focusing the lidar beams on that range. However, it is generally true that DIAL systems are best operated from an aircraft pointing down (e.g., Ismail and Browell, 1989), where the increase in backscatter with range counteracts attenuation due to absorption, rather than from the ground pointing up, which is the situation we must consider here. The second problem arises because the technique is differential in range. The noise level in the column content measurements is amplified when the measurements are differenced to obtain range-resolved estimates. This must be countered by use of some form of range-smoothing. The technique used here generates range-resolved difference estimates while at the same time smoothing the column content values by an amount that adapts to both the level of noise and the variability in the estimated absorption (section 5.6). Apart from making our results dependent on the validity of our smoothing technique, this raises a question of how to interpret the 'range resolution' of the estimates, which here will depend not only on how frequently the return time series is sampled, the rangegate, and the duration of the lidar pulse (which are the contributions usually considered), but also on the smoothing introduced. Fortunately the smoothing algorithm we use generates a 'smoothing distance' that is easily interpreted. This is what we shall quote in place of apparent range resolution, but it should be appreciated that this is a very conservative measure. A better technique of analyzing range resolution is being developed for DIAL analysis which will result in values more indicative of the true range resolution achieved by this method. This will be included in the final report. As with all optical systems, all-weather operation of DIAL is not possible because dense cloud cannot be penetrated.

The differential nature of DIAL does give rise to certain compensating advantages. Absorption interference (differential absorption arising from species other than water vapor leading to bias in the water vapor estimate) only matters if the interferant is at range of the measurement - high levels of interferant at low altitudes do not affect high altitude estimates. In practice interference is unlikely to be a problem with the techniques considered, especially when the continuously tunable 2 μm sources are used. Likewise, use of a differential technique removes the need for calibration.

1.4 Layout of this report

In Part I, simulations are used to compare the likely performance of direct detection and heterodyne lidars using what are essentially currently available infrared lasers operating at 2 μm and 10 μm. Absorption data (absorption coefficients and their temperature and pressure dependence) are taken from the HITRAN data base. In the simulations shown here, atmospheric parameters have been based on the values specified in the US standard atmosphere model. Water vapor concentrations have been randomized about these values to add realism and make the processing problem more challenging. Useful range-resolved estimates cannot usually be obtained simply by differencing column content values because of the noise level, so that the range-smoothing and differencing operations needed for both column content and range-resolved estimates are combined using a
Kalman filter algorithm. This also provides as an output an estimate of the standard deviation of the water vapor measurement. The backscatter model is one developed previously. In addition, there is an extensive discussion on practical problems relating to the lidar systems primarily discussed, which are currently being developed for various applications at NOAA-ETL, with special regard to their capability for turn-key operation.

The general properties of the DIAL technique are introduced with the help of examples using simulations in the next section, section 2. The capability for obtaining vertical profiles of water vapor is discussed in section 3. Practical problems associated with the infrared systems modeled are discussed in section 4, with particular emphasis on their potential for turn-key operation.

The principles underlying the optical heterodyne DIAL system and the Kalman smoothing technique, much of which either does not appear in the literature or has not been previously collected into a single document, are described in appendices. The differences between direct detection and heterodyning that affect measurement of return power, and the theory underlying heterodyne lidar and the DIAL technique, are elaborated in the first appendix, which is numbered as section 5. Details of the theory and parameters used in these simulations are given in the second appendix, section 6. Thus equations have been largely relegated to these appendices. Conclusions are briefly discussed in section 7. References for the entire text are listed in section 8.

2 General properties of the DIAL technique

In this section we illustrate our discussion of the DIAL technique using simulations of three different systems applied to the same problem, that of measuring water vapor concentration along a nearly horizontal line of sight over a maximum range of 6 km. The simulated systems employ lasers operating at wavelengths of about 2 \( \mu \)m and 10 \( \mu \)m and use direct detection and heterodyne receivers.

Properties of these systems and parameters of the measurements are summarized in Table 1. The lidars are based on laser sources that are close to the current state of the art. For the heterodyne systems low pulse energies but high pulse repetition frequencies are required. At each wavelength the pulse energy is assumed to be 5 mJ. Available PRFs are about 1 kHz for the \( \text{CO}_2 \) laser and 200 Hz for the 2 \( \mu \)m source, so that the 1000 pulses assumed initially to be transmitted at each wavelength by the \( \text{CO}_2 \) laser would take only about 1 second whereas the 2000 pulses assumed for the 2 \( \mu \)m laser would take about 10 seconds. If the noise in a direct detection system arises only within the receiver (which we assume below), then the performance of direct detection systems depends only on the receiver aperture-transmitted energy product, i.e., on the overall transmitted energy, irrespective of whether this is provided in a sequence of pulses or in a single, high power, pulse. If however other noise sources such as background light and signal-generated noise are significant, then the single high-power pulse is preferable. Thus for direct detection, high pulse energy lasers that are proven and operate at low PRF are listed. In practice the \( \text{CO}_2 \) source for direct detection lidar might be a transversely-excited, atmospheric pressure, (TEA) laser, and the 2 \( \mu \)m source a flashlamp-pumped (rather than a solid state diode-pumped) laser. The most compact lidar of those assumed would be the heterodyne 2 \( \mu \)m system. The solid state laser source for this system is somewhat smaller than
the CO₂ sources, and the antenna diameter (about 5") for the heterodyne system is the smallest of those considered. The total energy transmitted from the 2 μm systems listed in Table 1 (10 J) is also lower than from the larger 10 μm systems (50 J).

For the first set of simulations, all atmospheric parameters, including the water vapor concentration, are assumed to have values specified by the US standard atmosphere (e.g., Anderson et al., 1986). As mentioned in section 1.2, the primary figure of merit for DIAL is the SNR, and values for the SNR expected of the four systems considered are shown in Figs. 1 and 2. The latter also shows the (single pulse) carrier-to-noise ratio for the heterodyne systems. The SNR of the direct detection lidars reaches very high values at short range but falls away rapidly, that for the CO₂ lidar becoming less than 10 at a range less than 2 km. By contrast, the heterodyne systems have SNRs that are only weakly dependent on range, because the principal contributor to the noise at all ranges is from speckle which is averaged by use of the high PRF.

A further advantage of the heterodyne systems is illustrated in Figs. 3 and 4 which give examples of 2 μm lidar return signals at the point in the receiver where they would be digitized. The dynamic range of the direct detection returns is at least 5 orders of magnitude, whereas that of the heterodyne system is less than 2 orders of magnitude. The reasons for this are the relatively weak dependence of the mean return signals in heterodyne lidar, as indicated by the carrier-to-noise ratio (CNR, see section 5.2) plots of Fig. 2, and because in heterodyning the signal amplitude is digitized before squaring. The high dynamic range of direct detection systems has to be reduced in practice by tailoring the overlap of the transmitted beam and the receiver field of view, and/or by programming the gain of the photodetector to increase its output during the interval in which the return is received.

Table 1: Properties of lidar systems simulated in this section. The lidar antennas are assumed to be 10 m above the ground and the line of sight is inclined at 10° above the horizontal.

<table>
<thead>
<tr>
<th>LASER</th>
<th>CO₂</th>
<th>2 μm</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAVE NUMBERS</td>
<td>974.6219, 975.9304 cm⁻¹</td>
<td>4943.887, 4948.046 cm⁻¹</td>
</tr>
<tr>
<td>RANGE GATE</td>
<td>60 m</td>
<td>60 m</td>
</tr>
<tr>
<td>NO. OF RANGE GATES</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>RECEIVER</td>
<td>heterodyne</td>
<td>direct</td>
</tr>
<tr>
<td>PULSE ENERGY</td>
<td>5.0 mJ</td>
<td>500 mJ</td>
</tr>
<tr>
<td>NO. OF PULSES</td>
<td>1000/10000</td>
<td>100</td>
</tr>
<tr>
<td>TYPICAL PRF OF LASER</td>
<td>1 kHz</td>
<td>10 Hz</td>
</tr>
<tr>
<td>RECEIVER DIAMETER</td>
<td>0.3 m</td>
<td>1.0 m</td>
</tr>
<tr>
<td>RECEIVER EFFICIENCY</td>
<td>8 %</td>
<td>10 %</td>
</tr>
<tr>
<td>SAMPLING FREQUENCY</td>
<td>2.5 MHz</td>
<td>2.5 MHz</td>
</tr>
<tr>
<td>PHOTODETECTOR NOISE EQUIVALENT POWER</td>
<td>-</td>
<td>6 x 10⁻¹³ W/Hz⁰⁻¹</td>
</tr>
</tbody>
</table>
Figure 1: Multi-pulse signal-to-noise ratio (SNR) of CO$_2$ (heavy solid lines) and 2 µm (solid lines) direct detection lidars for both reference and absorbed beams.

Figure 2: Multi-pulse SNR (upper pair of lines) and single pulse carrier-to-noise ratio (CNR) for CO$_2$ (heavy solid lines) and 2 µm (solid lines) heterodyne lidars; reference beams only.

Figure 3: Examples of single pulse returns for both absorbed and reference beams of direct detection 2 µm lidar.

Figure 4: Examples of single pulse return for both absorbed and reference beams of 2 µm heterodyne lidar. Parameter shown is amplitude or envelope of signal, after photodetector but before squaring.
Figs. 5 and 6 show examples of estimates of the column content of water vapor obtained using the simulated CO₂ lidar data. The log ratio of the averaged return power estimates at the two transmitted wavelengths is formed and the water vapor concentration is fitted to these log ratios. It is assumed here that absorbers other than water vapor and atmospheric parameters such as temperature have values given by the US standard atmosphere, so that the only cause for the log ratio to be other than one is the presence of water vapor. The effect of the low SNRs shown in Fig. 1 is that column content estimates obtained from 'raw' log ratios obtained directly without range smoothing are very noisy. These estimates even appear to diminish with increasing range which, of course, the true column content cannot. Better results at long range are obtained from heterodyne (Fig. 6) than from direct detection data (Fig. 5). The effect of smoothing using a Kalman filter (section 5.6) with the range-resolved absorption modeled by a random walk is also shown. This filter is adaptive and its gain is determined from a fit to the data. From this gain the smoothing distance (the distance over which the filter is effectively averaging the column content data) can be determined, and this is shown in Figs. 7 and 8 for all the lidars. The distance becomes large and unstable where the filter is not working very well. For the direct detection lidars this happens at ranges greater than 2 km, but the heterodyne lidars provide useful data to twice this value. It might be noted that the filters use data at the beginning of the return time series to settle down and that this gives rise to the large excursions of the smoothing distance at very short range for the heterodyne lidars. This is not a fundamental defect in the method. Though not pursued in this analysis, use of a shorter data time series will result in range resolution comparable to that indicated at longer ranges.

Range-resolved estimates of water vapor concentrations, obtained as the point-to-point differences of the column content estimates, are shown for all four lidars in Figs. 9-12. Values obtained from raw log ratios are very unstable and essentially worthless so that only estimates obtained from the Kalman filter algorithms are shown. This filter and smoother also provide an output an estimate of the standard deviation σ of the water vapor estimate and in these figures is shown the ±1σ range around the concentration estimates. These values are usually reliable but it is seen from the long-range data of Figs. 9-11 they cannot be relied on when filter ceases to be stable, as evidenced for example from the range-smoothing distances of Figs. 7 and 8. In general, useful results appear to be generated by the 2 μm direct detection system and the 10 μm heterodyne lidar, and excellent results by the 2 μm heterodyne lidar (which averages for 10 times longer than the CO₂ system, Table 1). The direct detection CO₂ system, which as it happens is the only one of the four types of system considered here that has actually been constructed for application to DIAL measurements (at Stanford Research Institute, for chemical agent monitoring), does not give good data using aerosol backscatter beyond about 1 km.

The expected standard deviation of estimates is usually determined in simulation studies by averaging the results of repeated experiments using different noise sequences, assuming that the water vapor concentration varies smoothly with range as in the examples given so far. Now, the performance of a range-resolved DIAL system depends critically on range-smoothing, because the final results depend on differencing data from successive range gates. In practice the limitation on range-smoothing is set by the variability of the data as well as on the noise sequence. The variability of the data depends in part on the variability of the water vapor absorption that is being estimated. Thus reliance on data obtained using atmospheric models in which the water vapor concentration
Figure 5: Column content of water vapor from simulations of direct detection 10 µm lidar. True value (US standard atmosphere): heavy solid line. Estimate from raw (but time averaged) data: dotted line. Estimate after range smoothing with Kalman filter algorithm: solid line.

Figure 6: Estimates of column content of water, vapor with same notation as in Fig. 5, for 10 µm heterodyne lidar.

Figure 7: Smoothing length from Kalman filter estimates of column content obtained for direct detection lidars - 10 µm (heavy solid line) and 2 µm (solid line).

Figure 8: As Figure 7 for heterodyne lidars.
Figure 9: Estimates of range-resolved water vapor content from direct detection 10 μm lidar. True value (US standard atmosphere): heavy solid line; estimate from Kalman smoother: solid line; estimate plus and minus estimated standard deviation: dotted lines.

Figure 10: Estimates of range-resolved water vapor content from 10 μm heterodyne lidar. Notation as in Fig. 9.

Figure 11: Estimates of range resolved water vapor content using simulations for direct detection 2 μm lidar. Notation as in Fig. 9.

Figure 12: Estimates of range-resolved water vapor content obtained for 2 μm heterodyne lidar. Notation as in Fig. 9.
varies smoothly can be misleading. In Figs. 13-16 are shown estimates for the four systems considered which are the same in all respects as in Figs. 9-12 except that (i) the water vapor concentration is assumed to vary considerably, by a factor greater than 2, and (ii) the integration times of the 2 μm and 10 μm heterodyne systems are now assumed to be the same, i.e., the 10 μm lidar is assumed to run 10,000 pulses instead of 1000, and both systems are assumed to integrate data over 10 secs (see Table 1).

The effect of these changes is to make the performance of the two heterodyne systems very comparable, but each is rather worse than that of the 2 μm system for the smooth water vapor concentration used for Figs. 1-12. The improvement obtained by smoothing can be seen by comparing Figs. 13 and 14 for the CO₂ system, in which the data of Fig. 13 is smoothed and that of Fig. 14 is not. In Fig. 13 the estimated standard deviation of the water concentration estimate is also shown, as is the range-smoothing distance, which increases at ranges greater than 4 km. Data for the 10 μm direct detection lidar is not shown as performance is unimpressive beyond the first kilometer. The range-smoothing distance is rather greater for the 2 μm heterodyne system which results in the estimate being somewhat oversmoothed, reducing the sensitivity to fluctuations in the water vapor concentration. However the standard deviation, and the error in the estimates, is usually less than 1 gm/m³ at ranges up to 4 km. The slightly inferior performance of the 2 μm system in this comparison in which the integration times are the same is probably attributable to its smaller optics, which leads to a slightly inferior single pulse CNR (Fig. 2). The performance of the direct detection 2 μm system is as expected very dependent on range, giving excellent precision below 1 km and following the fluctuations reasonably well to 3 km, beyond which it oversmoothes.

3 Vertical profiling of water vapor

3.1 Comparison with horizontal and near-horizontal profiling

The principal differences between vertical and near-horizontal profiling are (i) that backscatter from higher elevations can be very much less than from near the ground, (ii) that at higher altitudes the variability of atmospheric parameters such as backscatter and water vapor concentration is also likely to be very much less, and (iii) the round-trip attenuation of the lidar beams, especially the absorbed beam in a water vapor DIAL system, is generally reduced.

The effect of reduced backscatter is to reduce the CNR of the return and therefore to make it less likely that the condition for efficient processing of the return, CNR ≥ 1 (section 5.4), can be satisfied. This can be partially compensated by the reduced variability of the return, which should permit longer data integration times. For data taken above the boundary layer, integration times for other optical monitoring techniques are often of order 1000 secs (about 15 mins.) or longer, rather than about 1 - 10 secs considered in order to give reasonable time resolution in the preceding section. The use of larger telescope optics in heterodyne systems designed for high altitude monitoring with slow, or zero, scan speed can further improve the quality of the return. In heterodyne lidar there is the further ability to focus the beam (section 5.3) from a sufficiently large telescope on the altitude of interest,
Figure 13: Range-resolved water vapor content obtained from heterodyne CO$_2$ lidar (solid line) using Kalman smoother compared with true value (heavy solid line) which varies randomly in range by a factor of about 2. Also shown are ±1σ points (dotted lines) and range smoothing distance (solid line, righthand axis) from filter output.

Figure 14: From same data as Fig. 3, showing estimates of range-resolved water vapor concentration obtained by differencing raw column content estimates without use of smoother.

Figure 15: Estimates of water vapor content and range-smoothing length as in Fig. 13 obtained from simulations of 2 μm heterodyne lidar.

Figure 16: Estimates of water vapor content and range-smoothing length as in Fig. 13 obtained from simulations of 2 μm direct detection lidar.
or on the maximum altitude in a range-resolved system, to be exploited. The limitations on telescope
diameter are refractive turbulence (section 5.3.3) and the cost of large diffraction-limited optics.

Another possibility, which is not considered here, is to use the backscatter from cirrus cloud
which can exceed that from aerosol by several orders of magnitude and might be used with a high
PRF lidar (for rapid integration of returns from a passing cloud) to generate column content values
to a discrete set of high altitudes. Cirrus backscatter is being studied in a separate program at
NOAA-ETL for diagnostics of the cirrus clouds themselves using the CO₂ lidar.

3.2 CO₂ heterodyne lidar simulations

The graphs in this subsection show results of processing data obtained from simulations of CO₂
heterodyne lidars assumed to be pointing vertically upwards. For the first three, Figs. 17 - 18(b), the
water vapor concentration was assumed to be given by the model of the US standard atmosphere
resulting in a smoothly varying concentration that the filters should be able to follow fairly easily.

The smallest system considered is as in the preceding section (Table 1), with a 5 mJ beam
transmitted from a transmitter/receiver telescope of diameter 0.3 m with a collimated output. Here
it is assumed that beams at each wavelength generate 10⁴ pulses, i.e., that each transmit for 10
seconds. Satisfactory results, albeit with some evidence of oscillation in the output of the filters, are
obtained at altitudes up to about 3 km, but beyond this the outputs obtained for this system always
become unsatisfactory. Evidence for this in Fig. 17 is the rapid increase and the instability in the
smoothing distance, and an increase in the estimated fractional standard deviation beyond 1 gm/m³.
If data is processed above this altitude what typically happens is that estimates of water vapor
concentration also become unstable and frequently negative. Results can be improved by use of a
larger transmit/receive telescope (diameter 1 m) that can be focused at 5 km, which is the maximum
range from which we have here found useful data to be obtainable. In Figs. 18 are shown estimates
obtained for the larger system, processing either only the high altitude data [Fig. 18(a)] or the whole
data set [Fig. 18(b)]. The effect of focusing is clearly seen in both cases from the reduction in the
filter smoothing length around the focusing range. Somewhat better results are obtained by
processing the entire data set in this instance, probably because the high backscatter at short range
help to stabilize the adaptive processing filters when the measured concentration varies smoothly.
This is by no means always the case - indeed it is perhaps the exception to the rule.

The performance of the 0.3 m telescope system against a randomized water vapor concentration
shown in Fig. 19 is very comparable to that against the smooth concentration shown in Fig. 17. The
higher spatial frequency fluctuations in concentration are not followed but the larger swings and the
trends are, with a standard deviation of less than 1 gm/m³. The range limit of this system is again
about 3 km. Fig. 20 suggests that the same is not true of the system with mirror diameter of 1 m.
Only the trend is followed and this not very satisfactorily, despite the focusing at 5 km. Significant
improvement is obtained here only by increasing the supposed integration to 10⁶ pulses (about 15 min
at a PRF of 1 kHz), corresponding to an assumed fractional standard deviation in the return power
measurements of about 0.2%. It remains to be seen whether this is realistic in practice, but even so
Figure 17: Example of estimate of vertical profile of water vapor concentration (solid line, ±1σ points shown as dashed lines) obtained in simulation of 5 mJ CO₂ heterodyne lidar with 30 cm diameter optics focused on a range of 5 km and after 10 000 pulses. The true value (US standard atmosphere model) is also shown (heavy solid line) as is the smoothing distance used by the filter (dotted line, right hand axis). Only the data from altitudes below 3.5 km are processed.

Figure 18(a): As Fig. 17, except that telescope diameter is 1 m. Only data from above an altitude of 2.5 km are processed.

Figure 18(b): As Fig. 18(a), except that data from all altitudes are processed.
Figure 19: Simulated performance of the lidar of Fig. 17 (CO₂ heterodyne system with 30 cm telescope, focused at a range of 5 km and running 10 000 pulses) for a randomized water vapor profile.

Figure 20: As Fig. 19 except that the telescope has a diameter of 1 m and only data from above 3 km are processed.

Figure 21: As preceding Fig. 20 except that integration over 10⁶ pulses is assumed.

Figure 22: As for preceding figure integration of data from 10⁶ pulses is assumed; here the laser energy is also assumed to be increased to 20 mJ.
the 5 mJ system is still only capable of following the trend in the water concentration (Fig. 21). If in addition the transmitted energy is increased to 20 mJ from 5 mJ (or, equivalently, the backscatter is four times greater than assumed in our model), the water vapor variability is fully followed, with a standard deviation of only about 0.1 gm/m³ and a range smoothing distance at its minimum value of about 100 m. The dramatic improvement in the fit of the smoothing model for only a fourfold increase in energy that is evident in comparison of Figs. 21 and 22 suggests that these parameters (5 - 20 mJ per pulse, 10⁶ pulses, 1 m optics) are at the threshold of what is needed for measurements with this system at these altitudes.

3.3 2 μm lidar simulations - heterodyning and direct detection

Results from simulations assuming a 2 μm heterodyne lidar rather closely parallel those of the previous section. The small system (5 mJ and 0.125 m telescope integrating 2000 pulses over 10 seconds at 200 Hz) permits following the trend of a randomly varying water vapor concentration up to an altitude of about 3 km (Fig. 23, cf. Fig. 19). It is assumed for this data that advantage is taken of the focusing capability of the 2 μm system and the optics are focused at an altitude of 3 km. To obtain reasonable results from higher altitudes, larger mirrors, longer integration times and more laser energy are required, and the focusing distance is shifted to 5 km. For the simulations of Figs. 24 to 26 the telescope diameter is assumed to be 0.5 m. For Fig. 24, the number of pulses integrated is again 2000, but for Figs. 25 and 26 it is 200 000 (15 minutes integration at 200 Hz). For Figs. 24 and 25 the laser pulse energy is 5 mJ, but for Fig. 26 this is increased to 20 mJ. The results can be compared with those of Figs. 20 to 22, and the conclusions are similar. In each case the results for 2 μm are slightly inferior to those of the CO₂ system and this can probably be attributed to the smaller mirror areas at 2 μm, although as indicated in section 5.3 near field returns do not simply scale with mirror area because of spatial coherence effects.

Figs. 27 to 29 for the direct detection system can be compared with Fig. 23, Fig. 25, and Fig. 26 respectively. In all the direct detection systems a large telescope of diameter 1 m is assumed. For the data of Fig. 27, a total transmitted energy of 10 J (e.g., 2000 pulses of energy 5 mJ) is assumed. In order to extend the altitude of operation to 5 km, this is increased to 1 kJ (e.g., 5000 pulses of 200 mJ) for Fig. 28 and 4 kJ (5000 pulses of 800 mJ) for Fig. 29. The results for the heterodyne and direct detection systems, for a given total transmitted energy, are very comparable at 2 μm, but it should be remembered that in simulating the direct detection system, the possible influence of background light has been ignored.

3.4 Temperature dependence

In the results that have been presented so far, the parameters of the atmosphere used to fit differential absorption data have been assumed to be those of the US standard atmosphere, apart of course from the water vapor concentration that is being determined. In practice these parameters are either unknown or have to be determined from other measurements. Use of incorrect values leads to systematic errors in the water vapor concentration. The parameters that affect the concentration
Figure 23: Data from simulated 2 μm heterodyne lidar transmitting 2000 pulses of energy $5 \text{ mJ}$, using optics of diameter 0.125 m pointing vertically and focused at a range of 3 km. True water vapor concentration shown as heavy line, smoother estimate as solid line with ±1 σ estimates as dotted lines, and smoothing distance as dashed line (right hand axis).

Figure 24: As Fig. 23 except that optics diameter is 0.5 m and focus is at range (altitude) of 5 km. As can be seen, only data from altitudes above 3 km are processed.

Figure 25: As Fig. 24, except that 200 000 pulses of energy $5 \text{ mJ}$ are transmitted.

Figure 26: As Fig. 25, except that 200 000 pulses of energy $20 \text{ mJ}$ are transmitted.
Figure 27: Data from simulated 2 μm direct detection lidar transmitting 10J, using optics of diameter 1.0 m pointing vertically. Graph line styles are the same as in Fig. 23.

Figure 28: As Fig. 27, except that total energy of 1 kJ is transmitted, equivalent to, e.g., 5000 pulses of 200 mJ.

Figure 29: As Figs. 27 and 28, except that total energy of 4 kJ is transmitted, equivalent to, e.g., 5000 pulses of 800 mJ.
estimates include pressure and temperature, and of these the temperature dependence has the greater impact. To indicate errors caused by temperature uncertainty, we give in Figs. 30 and 31 respectively the range-resolved water vapor concentrations determined as before, assuming standard atmosphere values for temperature, in cases where the model atmosphere used to construct the data has temperatures 10 K above or below the standard atmosphere values. The 'small' heterodyne lidars that have been previously shown to provide useful estimates up to 3 km (e.g., Fig. 17) are used for this comparison. It is seen that for the absorption lines used here, temperature errors of this magnitude lead to errors in the water concentration of about 1 gm/m³, or more than 5 times the standard deviation of the estimate for the CO₂ heterodyne lidar, but that the temperature coefficients of absorption are lower at 2 μm so that the 2 μm system is significantly less sensitive to temperature.

Preliminary studies have been undertaken to determine whether DIAL can be usefully applied to temperature profiling. The technique of interest makes use of the CO₂ lidar to measure atmospheric CO₂ absorption, the on-resonance lines being obtained as before using the C¹₂O¹⁶ isotope and the off-resonance using a laser containing the C¹⁸O¹⁶ isotope. The temperature coefficient of the CO₂ absorption line is therefore used as the monitor. Initial results suggest that temperatures might be monitored with a precision of 2-3 K in this way, which would be sufficient to support the water vapor measurements described above, but the precision would probably not be high enough for the technique to be offered as an alternative to current methods of tropospheric temperature estimation.

The validity of these results depends on the absorption model being used (e.g., its pressure and temperature dependence) and the validity of the absorption coefficients that here are taken from the HITRAN tapes (e.g., Rothman et al., 1987, see section 6.5, Grant, 1990).

4 Turn-Key Technology Assessment

4.1 Introduction

Until recently, lidar systems have been difficult and expensive to operate and maintain, largely because they were built as research instruments, often employing immature technologies unavailable from commercial sources. A typical lidar operation would require one or more PhD's, or equivalent, to align and maintain critical components of the system on a frequent basis, and several technicians to sustain day to day operations. Equipment failures were frequent, particularly with lasers and computer systems. Eye-safety has also been an issue because high power lasers operating in non-eye-safe wavelength regions were deemed necessary to achieve viable measurements with available technologies. This mode of operations is unacceptably expensive, insufficiently reliable, and potentially unsafe for routine, continuous water vapor profile acquisition operations required by CART sites and other atmospheric monitoring programs. Laser and computer technologies have matured significantly over the last half-decade, rendering routine lidar operations more a matter of good engineering than tenacity, and making possible technologies considered impractical or impossible just a few years ago. We consider here some of these newer technologies which we feel are capable of being engineered into eye-safe, near turn-key water vapor DIAL systems suitable for extended deployment with low operational cost and high reliability.
Figure 30: Estimates of water vapor concentration obtained in simulations of 5 mJ CO$_2$ heterodyne lidar (0.3 m telescope, 10 000 pulses) assuming temperature to be that of US standard atmosphere when in fact the temperature is either 10 K greater (solid line) or 10 K less (dotted line). The heavy solid line gives the true value in either case, which has been set to the standard atmosphere value used in earlier examples.

Figure 31: As Fig. 30 for simulations of the 5 mJ 2 µm heterodyne lidar (0.125 m telescope, 2000 pulses, focused at 4 km)
The critical technology requirements for turn-key operation of DIAL systems are: absolute wavelength stability, alignment stability, detector linearity, low maintenance components, and freedom from frequent calibration. In this section, the status of these critical technology areas is discussed with regard to DIAL systems operating in the 2 μm and 10 μm wavelength regions modeled in this report.

4.2 2 μm Wavelength region

4.2.1 Background

Over the last few years, there has been a flurry of development in solid-state lasers suitable for lidar systems operating in the 2 μm wavelength region employing thulium- and/or holmium- doped YAG materials with many variants of both crystal and dopants. These lasers have been pumped both by flash lamps for high pulse energy but low PRF, and by laser diodes, which enable high repetition rates, smaller packages, and more reliable operation. Development of the latter technologies are being fueled largely by the prospects for compact, lightweight, range resolved wind sensors for military and civilian aircraft. To date, a handful of 2 μm wavelength coherent Doppler lidars have been built [notably by Coherent Technologies (Henderson, et. al., 1993) and NOAA (Grund, 1995)]. Although water vapor DIAL has been proposed (Kavaya, et. al., 1989), no DIAL systems have yet been built in this wavelength region. The existing systems have been optimized for Doppler frequency measurement rather than for the intensity measurement required for DIAL.

The NOAA 2 μm High Resolution Doppler Lidar (HRDL) was constructed to achieve high spatial and temporal resolution boundary winds to study fine scale structure and turbulent transport processes in the boundary layer. A consideration in the design of this system was the potential for future water vapor DIAL operation. What is reported here is largely based on the experience of the authors with the development of the NOAA 2 μm lidar system technology. Although this is somewhat anecdotal, it is probably more reliable than manufacturers' claims, and represents actual field experience.

4.2.2 Wavelength stability

Injection seeding is typically used to assure single frequency operation of pulse lasers operating at these wavelengths because of the limited gain of the available laser media. The wavelength stability of the pulse is essentially that of the injection laser. Practical DIAL measurements in this region require the transmitting laser wavelength to be known to better than 10% of the absorption line width or ~1 ppm. This can be achieved straightforwardly using wavemeters and active feedback loops. However, this technique requires frequent calibration and wavemeters require considerable maintenance. Robust, although complicated, techniques involving active wavelength locking to water vapor absorption cell transmission are also possible and have been demonstrated at other wavelengths.

Two years experience with a commercially available CW, diode-pumped, single-mode Tm:YAG laser oscillator used in the NOAA system (CLR Photonics model CLR-2) suggests active wavelength controls may be unnecessary. Precise wavelength control is important to the NOAA Doppler system
because maximum range depends largely on minimizing attenuation by avoiding operation on water vapor and carbon dioxide absorption lines. This continuous-wave (CW) laser employs a reentrant cavity and two tilt-tuned etalons with precise temperature control to set the operating wavelength. The wavelength has been checked with a Burleigh model WA-20 wavemeter many times over the last two years and found to be stable to its original setting to at least 2 ppm (no change in wavelength has been noted and this is the stated accuracy for the wavemeter) without readjustment, despite having been roughly moved on more than one occasion and operated in a field trailer for several months. Power output at 2 μm has been stable to within 5% for the same pump diode current over this time period as well. This experience suggests the technology is commercially available to achieve both the wavelength stability and the reliability criteria for a turn-key water vapor DIAL system. If additional wavelength stabilization were desired, this could easily be implemented by electronically altering the cavity temperature set point in the laser controller.

Though tunable, these CW lasers are not frequency-agile in their present form. To achieve two wavelength operation as required for DIAL measurements, separate identical CW seed / local oscillator lasers could be employed, tuned to the on-line and off-line frequencies. The laser outputs could simultaneously be injected into the host pulse laser cavity. On-line / off-line frequency hopping could be achieved on alternate pulses by changing the length of the piezo-electric transducer (PZT) supporting one of the mirrors in the host laser (i.e. changing the host laser resonance frequency) by a fixed amount to assure resonance with the appropriate seed laser. The reference resonance position (probably the on-line frequency) of the host laser would be maintained as is presently done in the NOAA system (Grund, 1995).

If coherent detection is employed, the receiver local oscillator wavelength must also be switched to correspond to the transmitted pulse wavelength. The local oscillator beam impinging on the detector can be smoothly changed without moving parts by using standard electro-optic switching techniques. More elegant techniques using a single monolithic non-planar ring laser, which can be electronically tuned by varying the pressure on the crystal exerted by a PZT, have been demonstrated by Lightwave Electronics. However, frequency stability of these lasers has not been studied, and they are of relatively low output power.

Of course, a brute force approach is also possible where essentially two separate laser transmitters are operated simultaneously. This is not desirable because of the significant cost of duplicate lasers, the additional potential maintenance, and the problems associated with aligning two systems with separate optical paths to cancel non-linearities in the range response function (see next section).

4.2.3 Alignment stability

Solid state lasers are extremely stable. Meandering of the pointing direction and focus of solid-state lasers is largely a function of thermal variations in the laser medium and has not been observed to be a problem for steady-state operations at the level necessary to support coherent detection lidars. Because of the inherently small size of these lasers (e.g., the NOAA pulse laser is only 350 mm x 150 mm. x 150 mm), it has proven relatively easy to assemble a stiff structure which will maintain alignment over long periods and laboratory temperature ranges. The commercial aircraft wind sheer sensor lidar designs are completely hands-free operations in more hostile environments. Initial
(unofficial) test reports suggest this goal is almost but not quite achieved for the aircraft sensors.

Intensity measurements using coherent detection pose additional stringent constraints on alignment because the return signal light must not only fall on the detector, but must match the local oscillator wavefront. For a DIAL system employing a single laser source as suggested in section 4.2.2, the additional alignment requirement should not effect measurement accuracy because the on and off line laser source and receiver paths are the same. Thus, because the measurement is derived from the ratio of the signal range derivatives, and systematic range dependencies unrelated to the measured species are identical for the on and off line returns in this configuration, the effects of alignment errors on water vapor retrieval accuracy are effectively cancelled. Indirectly, a misalignment will cause a decrease in signal to noise ratio, and consequently degrade maximum range performance, however, this is much less important than possible systematic shifts in water vapor concentration. The room-temperature-operable InGaAs PIN diode detectors commercially available for this wavelength are of high quality and very compact size facilitating solid mounting. Again, compactness of the laser sources and other components facilitates engineering lidar systems for robust alignment stability.

4.2.4 Detector linearity

Typical parameters for commercially available room temperature InGaAs detectors, manufactured by EPITAXX Inc. as used in the NOAA lidar, have been incorporated into the system models in this study. These detectors have proven capable of linear operation over >5 orders of magnitude when used for direct detection, however, coupling this signal into high speed amplifiers diminishes performance considerably. Casual observations in the laboratory suggest the signal correlated noise and non-linearities, often associated with photomultipliers and avalanche photodiodes at shorter wavelengths (e.g. afterpulsing and delayed signal-dependant changes in responsivity), do not seem to be present at significant levels, though a careful study needs to be done if these detectors are to be used for direct detection DIAL.

Coherent detection removes much of the concern over detector linearity associated with DIAL operations. As mentioned earlier, the coherent detection process measures the amplitude not the power of the return signal. Thus the dynamic range requirement for a coherent detection system is reduced to the square root of the return power dynamic range required by direct detection systems. Additionally, the detector is subject to a continuous local oscillator induced bias which overwhelms the actual return signal power. Thus, changes in return signal power only slightly change the detector operating point. This also simplifies matching the signal into the amplifier, allowing greater linear range in the signal electronics. Wide band shot noise exceeding other noise sources by more than 20 dB has been demonstrated with the detector on the HRDL system.

4.2.5 Maintenance and operational requirements

Based on the commercial success of diode-pumped solid state lasers at other wavelengths, it was assumed, when NOAA embarked on developing its 2 μm Doppler system, that diode-pumping of a Tm:YAG solid-state laser would be essentially a turn-key process with little or no maintenance requirements over many years of operation. This has not been our experience. A fiber-optically coupled pump module consisting of ten 3 W laser diodes, coupled to the fibers through eutectically
bonded optical elements that are hermetically sealed in a cooled box with common power supply and thermal management, was built for us by Lightwave Electronics as a serial #1 prototype for a commercial product. While this module was completely satisfactory when it arrived, it has since developed many serious defects which would render it unsuitable for turn-key operations. Most of these problems relate to the operation of the power supply electronics and, we believe, to misalignment of optics within the sealed module. Superior commercial alternatives exist at this time however. Spectra Diode Labs, and Opto-Power Inc., both diode laser manufacturers, recently announced products in which diode laser bars are coupled to fibers at the diode level, which deliver powers comparable to the current Tm:Lu$_3$YAG laser requirements, at one fourth the cost of the Lightwave unit. The Opto-Power device has also achieved higher brightness then currently available as well. This is a superior approach, and is likely to solve the reliability problems we have experienced.

4.2.6 Calibration requirements

Providing the same laser transmitter and receiver paths are maintained for both the on and off line lidar signals as outlined in section 4.2.2, no special calibration of the instrument should be required beyond proper alignment to ensure best signal to noise ratio. However, a DIAL measurement can only be as accurate as the knowledge of the absorption cross section of water vapor in this region allows. There is some question as to the accuracy of the currently available laboratory measurements, and this may require further study beyond the scope of this report.

In a system incorporating separate transmitters and/or receivers, care must be taken to assure the same (or at least known) receiver-transmitter overlap functions (for the incoherent case), or beam size and wavefront curvatures (for the coherent case), exist for the on and off line wavelength channels. Discrepancies here would lead to differences in the range dependant sensitivity of the two channels, and distort the water vapor retrievals. Consequently, the one-laser one-receiver approach is preferable for the design of turn-key systems using coherent or incoherent detection.

4.3 10 $\mu$m wavelength region

4.3.1 Background

Tunable CO$_2$ laser technology in the 9 - 11 $\mu$m region has been available for a number of years. Until recently, most CO$_2$ lasers suitable for aerosol backscatter heterodyne lidar used discharge or e-beam pumping and were injection seeded to obtain single frequency operation. These systems tended to be quite large and somewhat difficult to align, control, and maintain. They were designed to deliver high pulse energies at relatively low pulse repetition rates as needed for long range Doppler wind measurements, although attempts were made to use them for DIAL (Hardesty, 1982) and a system of this type has been constructed specifically for DIAL (Grant et al., 1987). Emerging CO$_2$ technologies employing CW and pulsed radio-frequency (RF) excitation in small bore laser cavities are becoming available which produce similar average powers but do so with lower energy pulses developed at high repetition rates. Lower power versions of these lasers have been used in autonomous airborne lidars, both for research and in military systems, for over a decade. The advantage to these laser designs are smaller size and weight, lower cost, longer life, and less costly
maintenance (owing to lower fluence on optical surfaces). This technology has made possible master oscillator power amplifier (MOPA) laser designs which facilitate tight wavelength control and spectral purity as well. Because of these advantages, NOAA ETL is developing a lidar system based on a compact, frequency agile, high repetition rate, RF excited, MOPA CO₂ laser (Pearson, 1993). The system is designed to be suitable for both Doppler wind velocity measurements and for DIAL measurement of many species including water vapor. Projected performance characteristics of this system have been used as the model parameterizations in this report, and they are outlined in Table 1. Much of what is reported in the following sections is based on experience gained in developing this lidar and the MOPA laser.

4.3.2 Wavelength stability

CO₂ lasers suitable for lidar systems operate only on discrete wavelengths near the CO fluorescence peaks. Frequency stability of CW CO₂ lasers is assured by servo locking the laser at, or within a constant offset from, these gain peaks. In the MOPA design, a CW single-frequency laser is tuned by an intracavity grating to one of the known CO₂ molecular fluorescence peaks. A small dither of the cavity resonance is imposed by longitudinally driving one of the cavity end mirrors with a varying voltage applied to the PZT supporting the mirror. The laser output is observed with a lock-in amplifier, and the synchronous peak optical output power is maintained. The NOAA MOPA employs two simultaneously operable CW lasers locked to on-line and off-line frequencies for DIAL operations. Pulse operation is obtained by acousto-optically chopping the CW beams and alternately passing them through two multi-pass laser amplifier stages. Thus, the transmitted pulse has the frequency stability characteristics of the CW laser, and sufficient energy to provide viable returns using heterodyne detection. This technology has been demonstrated in the laboratory. The laser lines are chosen once during setup with the aid of a commercial wavemeter. Thereafter, active control maintains operation near the peak of that line. Field experience is not yet available, but wavelength stability problems are not anticipated.

4.3.3 Alignment stability

Because the laser is operated in a quasi steady-state, thermal beam steering is not observed. The small bore of the laser cavity made possible by the RF excitation minimizes refractive index variations due to gas turbulence, and gain irregularities due to non-uniform relaxation processes. These effects are troublesome in larger, lower repetition rate discharge-pumped lasers. Preliminary observations do not indicate the presence of these destabilizing mechanisms.

To achieve the gain necessary to develop 5 mJ pulses, calculations suggest the first optical amplifier requires 12 passes, and the second amplifier requires 6 passes. Stability of the multi-pass amplifier alignments has not yet been established, but, the compactness of the design suggests few problems with engineering a robust stable structure suitable for turn-key operations. This has yet to be demonstrated.

4.3.4 Detector linearity

Because of detector thermal noise at these wavelengths, cooling the HgCdTe (or similar) signal detectors to liquid N₂ temperatures is necessary to achieve the required SNR. Responsivity is not as linear with signal intensity for these detectors as for the PIN diodes available at shorter
wavelengths, and adequacy for incoherent DIAL use has not been established. When used for heterodyne detection, our recent laboratory studies (in progress) suggest select commercial detectors are capable of 40 dB linear dynamic range. To accomplish this range, and to flatten the frequency response, the detectors must be operated in the photo-conductive mode (i.e. with reverse bias). With careful balance of reverse bias and local oscillator power, 13 dB of wide-band shot noise can be achieved near 10 MHz with 1.2 dB/octave rolloff of shot-noise with frequency. As previously noted, heterodyne detection has the additional advantage of being responsive to the amplitude rather than the power of the signal return.

Signal correlated noise has not been observed with heterodyne detection and is not expected for the reasons previously mentioned (see 4.2.4). Signal correlated noise characteristics of these devices used as incoherent detectors has not been established.

4.3.5 Maintenance and operational requirements

As with any gas laser, leaks and gas mixture contamination are considerations for maintenance cycles. Fortunately, no exotic materials, arcing potentials, or highly corrosive gasses are present in the RF excited CO₂ laser considered here. Several CW laser sources have been built and operated over extended periods. We do not know the mean time before failure for these lasers because we have not yet had a failure. To date, the longest running laser of this design has been in operation for several hundred hours. A commercial Synrad sealed laser tube with integral RF drivers, of a type designed for medical and industrial applications, is now employed for the CW laser media, and is placed within a cavity of our own design to achieve single frequency operation. This has worked well for the local oscillator and injection oscillators lasers in the NOAA high power CO₂ coherent Doppler system. Recent improvements in the design of the amplifiers for the MOPA system have solved former leak problems. High sensitivity helium leak tests indicate ~1x10⁹ year time constant for the leak rate of the current design. This result was not easily reproduced after disassembly and reassembly of the unit. With some practice, we believe insignificant leak rates can be achieved operationally. Sufficient data is not yet available to analyze the potential gas contamination rate during operation, but we have reason to believe it will be insignificant because of the absence of reactive materials inside the lasers and amplifiers.

Optical damage potential is very low in this system owing to the low pulse energies and the IR wavelength. Replacement of optics is not expected to be a routine maintenance item.

The need for regular replenishing of liquid nitrogen in the detector Dewars persists, but is expected to be only a minor maintenance issue at the CART site. Although detector housings containing Sterling-cycle coolers exist, they are quite expensive, and have, at best, ~few thousand hour lifetimes. Suitability of Sterling-cycle coolers for coherent detection systems is also questionable because they contain reciprocating mechanical components which may introduce detrimental vibrations which would be difficult to isolate in a practical system.

4.3.6 Calibration requirements

Because the same laser transmitter and receiver paths are used for both on and off line wavelengths, the same arguments hold for the RF excited CO₂ MOPA technology as for the 2 μm
technology outlined in section 4.2.6. Greater care must be taken in wavelength selection because of
the potential for interfering species absorptions, and because the laser may only operate on discrete
lines. This is achievable, however, because of the broad tuning range.

4.4 Data Acquisition Processor Technology

A turn-key water vapor DIAL system at a CART site would likely be operated in a near-
continuous mode. This suggests data must be processed and reduced to concentration profiles in real
time. The technology for digitizing, averaging, filtering, and processing incoherent lidar returns has
been adequate for this task for many years. Heterodyne detection, however, requires much higher
digitization speeds and considerably greater digital signal processing capability to make the necessary
signal averaging possible for the high repetition rate systems considered in this report. The required
technological capabilities have become available commercially only in the last few years, largely
driven by market demands for multi-media and communications equipment. Although still somewhat
expensive, the processing hardware necessary for routine real-time heterodyne DIAL operations with
the straw-systems evaluated in this report are available commercially, and the cost/performance is
presently declining rapidly.

5 Appendix 1: DIAL equations

5.1 Precision of DIAL estimates - the lidar equation

The principle underlying the DIAL technique is that the trace concentration of absorbing species
can be inferred from comparison of the return powers obtained from lidars operating at different
wavelengths. Lidar return power is given by the lidar equation, which for targets such as aerosol
particles distributed in range is

\[ P_R(\lambda, r, t) = \int_0^\infty \frac{A \beta T}{r^2} P_T \left( \lambda, 0, t - \frac{2r}{c} \right) dr \approx \frac{A \beta T}{2r^2} c E_T(\lambda) \]  

where \( A \) is the antenna area, \( \beta \) the volume backscatter coefficient, \( P_T(\lambda, r, t) \) the transmitted pulse
power at wavelength \( \lambda \), range \( r \), and time \( t \), and \( T \) the atmospheric attenuation over the round-trip
path which includes the absorption to be compared. The factor \( A \beta T/r^2 \) is a function of range, but to
obtain the last expression on the right-hand side of Eq. (1) its variation has been considered negligible
across the scattering volume. Only the transmitted power is then integrated to give the laser energy
\( E_T(\lambda) \) (the 'short pulse' approximation).

The absorption component of the atmospheric transmission loss can be described by the Bouguer-
Lambert-Beers law

\[ T(\lambda, r) = e^{-2\pi(\lambda, r)} \]  

18
where the dependence on range and wavelength has been made explicit, the factor 2 comes from the lidar round-trip, and \( \tau(\lambda, r) \) is the range-integrated optical thickness. It is convenient to divide the latter into two components, the first arising from the specie to be measured, subscripted \( i \), and the second from other species, subscripted \( e \). Then

\[
\tau(\lambda, r) = \tau_i(\lambda, r) + \tau_e(\lambda, r)
\]

with

\[
\tau_i = \frac{r}{\int_0^r \sigma_i(\lambda) \eta_i(r') \, dr'}
\]

where \( \sigma(\lambda) \) is the absorption cross-section and \( \eta_i(r) \) is the concentration of the specie to be determined. There are similar formulas for each component of \( \tau_e \).

In the simplest analysis of the technique measurements are done at only two wavelengths, \( \lambda \) and \( \lambda + \Delta \lambda \), and it can be assumed that all the above quantities are independent of wavelength apart from \( \tau_i(\lambda, r) \). For the return from a single range, substitution of Eqs. (2) - (4) into (1) results in the following expression for the range-integrated or column differential absorption

\[
\Delta_\lambda \tau_i(r) = \tau_i(\lambda + \Delta \lambda, r) - \tau_i(\lambda, r) = \frac{1}{2} \log_2 \left[ \frac{P_R(\lambda, r)}{P_R(\lambda + \Delta \lambda, r)} \right]
\]

A difficulty with measurements from a target such as aerosol that is extended in range is that returns at any instant \( t \) cannot be attributed to a single range but come from a spread of ranges. If it is assumed that the range increment \( r \) over which the absorption is to be measured is much greater than the spatial length of the pulse then the range-resolved absorption can be found by differencing two column absorptions to give

\[
\Delta^2_\lambda \tau_i(r) = \Delta_\lambda \tau_i(r + \Delta r) - \Delta_\lambda \tau_i(r) = \frac{1}{2} \log_2 \left[ \frac{P_R(\lambda, r)P_R(\lambda + \Delta \lambda, r + \Delta r)}{P_R(\lambda + \Delta \lambda, r)P_R(\lambda, r + \Delta r)} \right]
\]

This is doubly differential, over both wavelength and range, and like Eq. (5) it is nonlinear in the return powers \( P_R \).

If all the parameters of these equations are constant in time so that independent measurements of return power can be described in terms of ensemble averages \( \langle P_R(\lambda, r) \rangle \) and \( \langle P_R(\lambda + \Delta \lambda, r) \rangle \), together with a variance about these averages, then differentiation of Eq. (6) leads to
\[
\var\left[\Delta^2_{\lambda,r}\right] = \frac{1}{4} \left( \var\left[\log_e P_R(\lambda,r)\right] + \var\left[\log_e P_R(\lambda + \delta\lambda,r)\right]
+ \var\left[\log_e P_R(\lambda,r + \Delta r)\right] + \var\left[\log_e P_R(\lambda + \Delta\lambda,r + \Delta r)\right]\right)
\approx \frac{1}{4} \left( \nvar[P_R(\lambda,r)] + \nvar[P_R(\lambda + \delta\lambda,r)]
+ \nvar[P_R(\lambda,r + \Delta r)] + \nvar[P_R(\lambda + \Delta\lambda,r + \Delta r)]\right)
\]
so that the variance of the absorption estimate is related to the fractional or normalized variance of the return powers defined by

\[
\text{nvar}[P_R(\lambda,r)] = \var[\log_e P_R(\lambda,r)] \approx \frac{\var[P_R(\lambda,r)]}{<P_R(\lambda,r)>^2}
\]

The quantity needed to determine the precision of DIAL estimates is therefore the fractional standard deviation \(\text{fsd}(P_R) = \sqrt{\text{nvar}(P_R)}\). Values can rarely be expected to be below 1% and the corresponding limit on the precision of absorption measurements is easily calculated. Calculation of \(\text{fsd}(P_R)\) is the subject of the remainder of this appendix.

### 5.2 Direct detection and heterodyne lidar

The difference between heterodyne and direct detection lidar lies in the way the photodetector is used, which is summarized graphically in Fig. A.1. The photodetector is a square law device that gives an output current proportional to the square of its input, which is the amplitude of the optical signal. For direct detection, the mean output current is

\[
<i_{dd}> = \left(\frac{\eta e}{h \nu}\right) \int_{A_d} <u_s(s)> d^2s = R_s P_s
\]

where \(\eta\) is the quantum efficiency of the photodetector, \(\nu\) the frequency of the optical signal, and \(\eta e/(h \nu) = R_s\) is the current responsivity of the photodetector. The optical amplitude over the surface of the photodetector is described in (9) by the function \(u_s(s)\), which is normalized so that the integral in (1) equals the average optical signal power \(P_s\). Eq. (1) represents the action of a photodetector in direct detection as conversion of a stream of photons (power \(P_s\)) into a stream of photocarriers (current \(i_{dd}\)).

For heterodyning (again refering to Fig. A.1) the amplitude of the optical signal is combined with that of a monochromatic local oscillator (LO) beam of power \(P_L\), so that the total field is the sum \(u_{tot}(s) = u_s(s) + u_L(s)\), and this is squared across the surface of the photodetector to give
Figure A.1: Schematic diagram of direct and heterodyne detection receivers

Figure A.2: Schematic diagram illustrating addition of signals from $m$ coherence patches on photodetector
The first two terms in the final expression describe direct detection of respectively the signal and LO fields. In practice, a high-pass filter is used to remove these currents. The final term is the average of the heterodyne current \( i_{\text{het}} = 2R\mu_s(s)u_L(s) \) and represents the response of the photodetector to the moving interference pattern generated by the two fields. If the optical signal as well as the local oscillator is monochromatic and if each was spatially uniform, then \( i_{\text{het}} = 2|\mu_s||u_L|\cos(\omega t) \) and would oscillate at the frequency \( \omega = \omega_s - \omega_L \), which is the difference between the frequencies of the two input fields \( u_s \) and \( u_L \). In general, the optical carrier frequency in \( u_s \) is removed by the mixing process at the photodetector and its other spectral properties (mainly bandwidth, Doppler shift, and power) can be measured at radio rather than optical frequencies. However, because \( i_{\text{het}} \) is a.c. its average as written in (10) is in fact zero. To determine the mean power of the signal \( i_{\text{het}} \) has to be rectified before averaging. If a square law detector is used (either in hardware or, more usually, in software), we obtain for the mean current at its output

\[
<i_{\text{het}}^2> = \left( \frac{\hbar c}{\hbar \nu} \right)^2 \int_{A_d} \left( |u_s(s)|^2 + |u_L(s)|^2 + 2<|u_s(s)||u_L^*(s)|> \right) d^2s
\]

\[
= R_s(P_s + P_L) + 2R_i \int_{A_d} <u_s(s)u_L^*(s)> d^2s
\]

(10)

A factor 2 has disappeared because the average of \( \cos^2(\omega t) \) is 1/2. The spatial variation of the interference pattern across the surface of the photodetector (usually referred to as the spatial mismatch of the signal and LO fields) also leads to a reduction of the final output compared with what might be expected from the direct detection equation (9); this is taken account of here by introducing the received power \( P_R = \eta_{\text{het}}P_s \) where the reduction factor \( \eta_{\text{het}} \) is known as the heterodyne efficiency.

This introductory discussion highlights the following problems with heterodyning systems:

(i) A local oscillator laser is required operating at a single frequency (single mode) that is locked to that of the transmitter.

(ii) The electrical bandwidth of a heterodyne receiver is identical to its optical bandwidth. In direct detection systems, the optical bandwidth is usually minimized using optical filters to reduce background light but it usually remains very much greater than the receiver bandwidth, which is minimized separately using a low-pass filter to match the range resolution of the lidar to reduce the electronic noise in the receiver. In heterodyning, the optical and receiver bandwidths are the same, and are determined by the optical properties of the return such as its maximum Doppler shift and its bandwidth. Heterodyne receivers typically therefore have higher noise bandwidth and require high digitization and data rates, especially if operating at short wavelengths.

(iii) The heterodyne output is reduced by the factor \( \eta_{\text{het}}^2 \). Optimization requires, in addition to axial alignment, that the angular spread of the LO and the signal must be as small as possible,
resulting in a requirement for single spatial mode optics and diffraction-limited beams (section 5.3). Even so, the theoretical optimal value of $\eta_{\text{opt}}$ for targets giving rise to (spatially) incoherent backscatter is less than 50% (Rye and Frehlich, 1992). The return signal is further reduced by refractive turbulence in the optical path which increases the angular spread of the transmitted beam and the field of view of the receiver.

(iv) Fluctuations in the optical signal $u_0(s)$ appear as modulation on the a.c. component of the photodetector output ($i_{\text{det}}$) and cannot be averaged until after rectification [in direct detection fluctuations in $|u_0(s)|^2$ are integrated over the photodetector surface, as is described by the integral in Eq. (9) - which is known as aperture averaging]. Spatial fluctuations in lidar signals are implicit in returns from rough or diffuse targets (spatial speckle). Temporal fluctuations also occur with bandwidths about equal to that of the return (coherent fading, or temporal speckle) and, with lower bandwidths, caused by fluctuations in $\beta$, $\tau$, and as a result of atmospheric refractive turbulence (scintillation).

(v) Rectification after coherent addition of the electrical signal and noise in the output of the photodetector degrades the final signal to noise ratio when the mean signal level is small compared with the root mean square (rms) noise, as a result of mixing in the rectifier (section 5.4).

The question has to be answered - why in the face of these problems should heterodyning be used at all? The principal advantages of heterodyne lidar are:

(a) After the squarer, the mean signal or carrier current $<C> \propto <|i_{\text{det}}|^2>$, as given in Eq. (3), whereas the root mean square (rms) of the squared photodetector noise

$$<N> \propto <|i_M|^2> = R_i^2 P_L [2 F_Q \eta v B]$$

where $F_Q$ is a quantum noise figure that is the ratio of the actual noise level to the minimal value for photon-induced shot noise $\eta v B$. The carrier-to-noise ratio (CNR) is defined as the ratio

$$\text{CNR} = \frac{<C>}{<N>} = \frac{|i_{\text{det}}|^2}{|i_{\text{det}}|^2} = \frac{P_R}{F_Q \eta v B}$$

$F_Q$ is a figure of merit that has the minimum value of unity in the limit that the only noise source is LO-generated shot noise and the photodetector quantum efficiency is unity. The shot noise can be increased until it dominates other sources simply by increasing the LO power. This also amplifies the heterodyne signal so that the CNR becomes independent of $P_L$. If $F_Q = 1$ then $\text{CNR} = P_R/(\eta v B)$ which is the minimum value given the quantum nature of light. In practice $F_Q$ is of course somewhat larger than 1, but it is usually in the range from about 2 to 4. Heterodyning is therefore very sensitive to low level returns which cannot otherwise be accessed in wavelength regimes such as the infrared where photon-limited direct detection is not available.

(b) Despite this high sensitivity, the narrow field of view imposed by heterodyne efficiency optimization [(iii) above] and the relatively narrow optical bandwidth imposed by RF filters [(ii)] have
the consequence that heterodyne receivers are very insensitive to thermal background light. This ensures that daytime operation of heterodyne systems presents no special problems and is indeed routine.

(c) Compared with direct detection, the signal obtained from heterodyne receivers have a limited dynamic range, in part because of the range dependence of the heterodyne efficiency [section (5.3)] and in part because immediately after the photodetector the signal is proportional to the amplitude of the carrier + noise, and not to its square, the power. It is usual to digitize the heterodyne return at this point.

(d) Photodetector 'after-pulsing' is of no consequence in heterodyning, and nonlinearities induced in detectors and amplifiers are greatly diminished or eliminated.

(e) Advantage can be taken of the laser frequency control needed in a heterodyne DIAL system to position the wavelengths of on- and off-line beams from a continuously tunable laser (like the 2 μm sources) very close, so that differential backscatter, attenuation, and interference absorption are unimportant.

5.3 Heterodyne lidar antennas

5.3.1 Introduction

The terms in the lidar equation (1) that are available to the lidar system designer are the receiver area A and the transmitted pulse energy E. In a direct detection lidar A might be simply the physical area of the primary receiver dish, although in practice the return may be reduced by vignetting in the secondary optics of the receiver which can be regarded as, in effect, a reduction in A. If all the other parameters in the lidar equation are constant, then Eq. (1) shows that the received signal would fall rapidly as 1/r², resulting in a signal of large dynamic range. Vignetting is sometimes therefore introduced to enhance returns from long distances at the expense of those from close-in and reduce dynamic range. Calculations of the field-of-view 'overlap' functions produced by these geometries are in principle straightforward, if cumbersome, and are made using geometrical optics. In the operation of heterodyne (or 'coherent') systems, the receiver properties are determined in part by the local oscillator geometry so that, as well as physical vignetting of both the received signal and the local oscillator beam (e.g., by physical apertures), any mismatch in the profile of these beams can result in a loss of output signal. This can also be accounted for by a reduction in A. Thus when discussing heterodyning in this section, the heterodyne efficiency \( \eta_{hd} \) mentioned in section (5.2) is included within A, which must be calculated using wave optics. To some this calculation appears more complicated and the results appear counter-intuitive.

The problem for heterodyne DIAL systems using aerosol and molecular targets are that the latter are randomly positioned in space and light backscattered from them is spatially incoherent at its source. The optical field at the receiver therefore has amplitude and phase that are random both spatially and temporally. Both \( |\bar{I}_{in}| \) and \( |\bar{I}_{out}| \) (Eq. (11)) still contain, unsmoothed, fluctuations arising from the input optical field. Thus if the final power measurement is described in terms of the lidar
parameter $A$, then $A$ must be regarded as a fluctuating quantity having an average value $<A>$ and a variance. In this section we discuss the calculation of $<A>$.

Some results of a simple calculation of $A$ are given in the next subsection. It is useful to introduce it with a simple argument from Cummins and Swinney (1970) that further illustrates the difference in the output of heterodyne and direct detection receivers. Although the received optical field is random, it does contain spatial correlation. In the surface plane of the photodetector, the correlation distance is approximately the separation of points that can just resolve the light source. The reason for this is that the ability to resolve (say) two sources, or two parts of the same source, depends on the ability to distinguish the angles of arrival of light from the two sources, and in wave optics the spread in the angles of arrival is characterized by the spread in the phase of the light. Over distances small compared with the resolution distance, the phase of all light from the source can be regarded as essentially constant, i.e., the illumination is coherent. Over distances large compared with the resolution distance the phase varies randomly. Suppose that we imagine the photodetector surface to be divided into patches of dimension equal to the resolution distance (Fig. A.2). Within each patch then the light is regarded as coherent. If the heterodyne signal from the photodetector is rectified by a square-law detector and the output is integrated over some short time (short compared with the correlation or reciprocal bandwidth period of the return, which fluctuates temporally as well as spatially), then the resulting charge from a single patch is $<Q_{het}> = e^2 n_s n_L$ where the $n$ are photocounts. If the signals from $m$ different patches are added before squaring then the resultant will be found by addition of phasors of random phase representing each patch, and the mean square amplitude of the resultant will be $<Q_{het}> = e^2 mn_s n_L$, i.e., it increases as $m$, not as $m^2$. Writing the output due to the local oscillator alone as $N_L = mn_L$, we find that the output of the heterodyne receiver is $<Q_{het}> = e^2 n_s n_L$, i.e., it is a measure of the product of $N_L$ with $n_s$, not with of the signal photocount $N_s = mn_s$. The quantity $n_s$ we have introduced as the photocount per resolution or correlation interval, but it might be understood easier as the photocount per speckle; in quantum optics where the product of the spatial and temporal resolution intervals that enter the uncertainty principle define a 'mode' of the optical field, it is termed the photocount per spatial mode or the (spatial) count degeneracy of the light. The upshot is that heterodyning measures the return signal or photocount from a target that is too small to be resolved, but for larger receiver optics the signal estimate $<Q_{het}> / e^2 N_L$ saturates and heterodyning gives the degeneracy. (This is the reason why heterodyne receivers are very insensitive to most thermal light, for which the degeneracy - which is equivalent to the occupation number of an energy state as given by Planck's formula - is less than one). Heterodyne receivers are usually set up so that the photodetector is illuminated by only one spatial mode, corresponding roughly to a single Airy disk within the receiver optics, which is sufficient to maximize the output. The transmitter of a heterodyne lidar is likewise set up so that the target at the range of most interest (perhaps the longest available range) is of a size that is just resolved by the receiver. This implies that the transmitter dish should be of about the same size as the receiver because if the transmitter is the smaller then the illuminated area of the target can always be resolved by the receiver and light scattered from the edges of this area miss the photodetector, while to make the transmitter larger than the receiver (so that the minimum target size is below the
resolution limit of the receiver) is unnecessary. In practice heterodyne systems are usually designed so that a single mirror is shared by the transmitter and the receiver. It might be noted before leaving this simple-minded argument that heterodyne systems are ideally set up so that the limiting noise in the system is the shot noise of the local oscillator, so that in the absence of signal the receiver output would be \( \langle Q_N \rangle = e^2 \text{var}(N_I) = e^2 N_L \). The CNR [Eq. (13)] is then \( \langle Q_{\text{het}} \rangle / \langle Q_N \rangle = n_s \), which is also the degeneracy, and shows again that there is no advantage in observing more than a single mode of the return.

5.3.2 Gaussian beam approximation

The derivation of equations for the effective antenna area \( A \) of heterodyne systems are too complicated to be included here, although the approach is straightforward enough if the concepts of Fourier optics are used (Rye, 1982), but we shall quote the results for an important special case that has been used in the DIAL simulations quoted in this report. The profile of the optical field at the transmitter antenna is assumed Gaussian

\[
|u_T(b)| = \sqrt{\frac{2}{\pi}} \left( \frac{P_T}{b_T} \right) e^{-b^2 b_T^2} \tag{14}
\]

where \( b_T \) is the \( 1/e^2 \) or irradiance radius of the beam (the geometry is sketched in Fig. A.3). The properties of the receiver are defined by an imaginary beam propagated back from the photodetector as described by the antenna theorem (Siegman, 1966), and its profile at the receiver antenna is assumed the same as (1), with the radius \( b_R \) replacing \( b_T \) and with the local oscillator power \( P \) replacing \( P_T \). It is assumed that there are no hard apertures in the system, in particular that the Gaussian beam profiles are not truncated at the edges of mirrors or at the photodetector. In the absence of refractive turbulence the lidar equation term \( <A> \) is then

\[
\frac{1}{<A>} = \frac{1}{A_T'} + \frac{1}{A_R'} \tag{15}
\]

where

\[
A_T' = \frac{2\pi b_T^2}{1 + m_T^2}, \quad m_T = \frac{\pi b_T^2}{\lambda} \left( \frac{1}{r_T} - \frac{1}{r} \right) \tag{16}
\]

\( r_T \) being the range on which the transmitter beam is converging and \( m_T \) a generalized Fresnel number. \( m_T \) can be interpreted in several ways. Consider transmission of a collimated beam, \( 1/r_T = 0 \). The angle subtended by the transmitter at any target range \( r \) is (about) \( b_T/r \). In Fourier optics terms, the spatial frequency bandwidth of the transmitter is then \( b_T/(\lambda r) \). \( m_T \sim b_T^2/(\lambda r) \) is therefore the size-bandwidth product of the transmitter as seen from the target, and \( m_T^2 \) the two-dimensional area-bandwidth product. In terms of traditional wave optics, \( m_T \) is the number of Fresnel zones within the transmitter as seen from the target. In the context of lidar \( m_T^2 \) is interpreted as the number of coherence areas or speckles in the light backscattered from the target that fit within the transmitter area. \( m_T \) is non-integer and may be less than one. \( 1 + m_T^2 \) has the property that it cannot be less than 1 and can be interpreted as the number of speckles 'seen' by the transmitter (the transmitter cannot
see less than a single speckle even if \( m_T \ll 1 \), and if \( m_T \gg 1 \), then \( 1 + m_T^2 \approx m_T^2 \). In the far field, \( 1/r_T - 1/r = 0 \), \( m_T = 0 \), and \( A_T' \) is maximized. \( A_T' \) might be called the effective area of the transmitter and is related in radar to the antenna gain \( G_T = A_T'/(\lambda^2/(4\pi)) \) (Rye, 1982). There is a similar expression to (16) for \( A_R' \) in which the subscript \( T \) is replaced by \( R \). Eq. (15) shows that \( \langle A \rangle \) is always less than the smaller of \( A_T' \) and \( A_R' \). Transmitter and receiver are coupled in the general expression for \( \langle A \rangle \), which is an integral quantifying the overlap in their fields of view, but for Gaussian beams they happen to be separable as shown in Eq. (15). In practice it is usually the case that the transmitter and receiver are matched both in size (as discussed at the end of the previous section) and in beam curvature, so that \( A_T' = A_R' = A_{TR}' \).

5.3.3 Refractive turbulence

The spatial correlation function for propagation of a beam through a medium exhibiting refractive turbulence fluctuations, perpendicular to the direction of propagation, is given by \( \exp[-(b/p)^{5/3}] \) where the transverse coherence length

\[
\rho = \left[ 1.45k^2 \int_0^r C_n^2(r') \left( 1 - \frac{r'}{r} \right)^{5/3} dr' \right]^{-3/5} \tag{17}
\]

with \( k = 2\pi/\lambda \) and \( C_n^2 \) the structure parameter for refractive index fluctuations. This may be brought into the Gaussian model if the exponent \( -(b/p)^{5/3} \) is approximated by \( -(b/p^2) \) (the 'quadratic' approximation, Yura, 1979) and defining the coherence area \( A_{turb}' = \pi \rho^2 \). If the path traversed by the transmitter beam lies outside the receiver field of view over most of the distance to the target (the 'bistatic' approximation) then the fluctuations induced by turbulence on the transmitted and backscattered light are independent and it can be shown that

\[
\frac{1}{\langle A \rangle} = 2 \left( \frac{1}{A_{TR}'} + \frac{1}{A_{turb}'} \right) \tag{18}
\]

This is the formula used in the simulations in this report. The factor 2 on the right hand side arises in both terms from the radar round trip. Turbulence has little effect if \( A_{turb} > A_{TR}' \). For a single transmitter/receiver (or transceiver), which is the usual arrangement in practice, neither the bistatic nor the quadratic approximation are applicable and the calculation is more complicated. Eq. (18) is however a good approximation in weak turbulence and is conservative, i.e. (Rye, 1981, pace Yura, 1979, Frehlich, 1993), for the transceiver geometry the value given by (18) sets a minimum for \( \langle A \rangle \). The factor by which \( \langle A \rangle \) in a transceiver system is enhanced is greatest in intermediate turbulence and is about 2 in strong turbulence. Consideration of the range-weighting in Eq. (17) shows that turbulence close to the transmitter or receiver has the strongest influence on \( \rho \) and \( A_{turb} \) (the 'shower curtain' effect, Yura, 1979).

5.3.4 Summary

The range dependence of the expected return calculated for a heterodyne lidar using these equations with atmospheric parameters that are constant in range and neglecting the path-length
Figure A.3: Optical beam profile and curvature

Figure A.4: CNR (lines) and SNR (dashed lines) of heterodyne lidars that are either focused at 2 km (heavy density print) or collimated (ordinary print), illustrating Eqs. (15)-(18)
effects of attenuation and refractive turbulence is shown in Fig. A.4 for transceiver beams that are focused and collimated. The former shows how focusing can be used to enhance the return signal from long range. In the latter $1/r_T = 0$ and the near-field return is more nearly constant because as long as $m_T$ and $m_R$ are greater than 1 the coherence areas $A_T'$ and $A_R'$ increase with $r^2$ and exactly compensate for the $1/r^2$ term in the lidar equation. This has the effect of diminishing the dynamic range of the heterodyne return as mentioned in section (5.2) under (c).

5.4 CNR, SNR, and fractional standard deviation

From the discussion in section (1.3), the mean signal in the heterodyne receiver after squaring (Fig. A.1) is $\langle C \rangle \propto P_L P_R$ and the mean noise level is $\langle N \rangle \propto P_L F_d \nu B$. The average output of the squarer is $\langle C + N \rangle \propto P_L (P_R + F_d \nu B)$. Because $\langle N \rangle$ is known (it can be determined when the signal is absent), an estimate of the signal power can be found by deducting $\langle N \rangle$ from an estimate of the output. To determine the fractional standard deviation of a measurement [as needed for Eq. (8)], the statistics of the combined signal and noise have to be considered. The statistics of both the signal $i_{s\text{et}}$ and the noise current $i_n$ after the photodetector are Gaussian, which will therefore also be the statistics of their sum. After squaring, the statistics of the output is exponential, giving for the second moment $\langle (C + N)^2 \rangle = 2 \langle C + N \rangle^2$ (the cross-term disappearing because $C$ and $N$ are independent). The variance of the output is $\text{var}(C + N) = \langle (C + N)^2 \rangle - \langle C + N \rangle^2 = \langle C + N \rangle^2$. If this is normalized to the square of the mean signal power estimate we find

$$\frac{1}{SNR} = \text{fsd}(P_R) = \sqrt{\frac{\text{var}(C + N)}{\langle C \rangle^2}} = 1 + \frac{1}{\text{CNR}}$$  \hspace{1cm} (19)

where $\text{CNR} = \langle C \rangle/\langle N \rangle$ (Eq. 13). The uncertainty in the estimate of carrier power comes from two sources, first the fluctuations in the signal itself (speckle), and second the receiver noise, arising from the local oscillator shot noise and from within the photodetector and associated electronics. The two terms on the right hand side of (19) can be associated respectively with these two components although it should be remembered that the signal and the noise are not separate in the receiver, having been added coherently at the output of the photodetector and in subsequent electronics. Techniques for the improvement of $\text{fsd}(P)$ by averaging depend on the bandwidth of the signal and noise which is discussed in the next section.

Fig. (A.4) includes plots of the variation of SNR with range that correspond to the plots mentioned in section 5.3.4 of CNR. The SNR can never be greater than one and if the CNR is less than one then SNR falls as rapidly as CNR. In general, heterodyne lidars are best designed to ensure that the CNR should be no less than one under normal operating conditions, so that only the speckle fluctuation has to be averaged. If this is done, there are two distinct reasons why the precision of heterodyne lidars should be rather constant with range. First, $\langle A \rangle$ and therefore the CNR has relatively weak range dependence for collimated beam geometry, and focusing can be used to emphasize the returns which might otherwise be weak (Fig. A.4); second, from Eq. (19), the SNR is a weak function of the CNR where the latter is greater than one.
The SNR of direct detection measurements is much more obvious, being

\[
(SNR)_{dir} = \frac{P_R}{(NEP)B_{dir}^{1/2}}
\]

where the noise-equivalent power NEP is by definition the signal power level that would give an output equal to the rms noise level at unit bandwidth (usually 1 Hz). The SNR here varies as the square root of the receiver bandwidth \(B_{dir}\) which is usually set to correspond to the range resolution required, \(B_{dir} = 2\Delta r/c\). \((SNR)_{dir}\) can be much larger than 1 but its dependence on range is the same as that of \(P_R\) as derived directly from the lidar equation Eq. (1).

If a direct detection receiver is signal photon-limited (which is not the case for the infrared systems considered here), then \(NEP = \sqrt{P_R h \nu}\) [c.f. Eq. (12)] and the SNR is proportional to \(\sqrt{P_R}\) not \(P_R\). This is a fundamental difference between 'quantum-limited' direct detection and heterodyne systems for which the SNR, using the simplified discussion of this section, is described by Eqs. (13) and (19) and is proportional at least for small signals to \(P_R\). Thus where quantum-limited detection can be used, direct detection has the advantages that for small signals the SNR does not deteriorate so rapidly as \(P_R\) is reduced, and for larger signals speckle can be aperture-averaged. The \(P_R\) dependence of the SNR of heterodyne systems can be regarded as a consequence of small signal suppression in the output of power law detectors (Davenport and Root, 1958).

### 5.5 Return power estimation in heterodyning

The precision of a lidar return power measurement is limited by the receiver noise. The noise level is determined by the receiver bandwidth which in a direct detection system is set by the range resolution, e.g., by the relation \(B_R = c/(2\Delta r)\), where \(\Delta r\) is the length of the range gate. In a heterodyne system the receiver bandwidth must be sufficient to accommodate all possible Doppler shifts; otherwise attenuation of the return can take place due to the signal drifting outside the reception band. The receiver bandwidth is therefore set at least equal to the Doppler search bandwidth \(F_s\), which is usually greater than \(B_R\). This would not matter if integration of the return over the range gate restored the original SNR, e.g., if the SNR improved as \(\sqrt{(F_s/B_R)}\), but as we shall see this is not usually the case either for averaging small signals out of background noise or for averaging speckle.

The time series return from the lidar can be characterized by its power spectrum or its correlation function. These functions are related as a Fourier transform pair but both are useful because it is sometimes easier to discuss a problem in terms of the one and sometimes in terms of the other. For calculations and simulations we shall use the following Gaussian functions. The power spectrum is

\[
P_s(f) = \sqrt{2\pi\alpha} \exp\left[-(F - F_1)^2/(2F_2^2)\right]
\]

where \(F_1\) is the Doppler frequency shift, \(F_2\) the return signal bandwidth and the (temporal) photocount degeneracy \(\alpha = N_p/m\), \(N_p\) being the average effective number of photocounts received and \(m = \)
$2\pi F_2 T$ a time-bandwidth product that gives the number of correlation intervals or 'speckles' that fit within the duration of the rangegate $T = 2\Delta r/c$. $m$ is non-integer and may be less than one. The corresponding (complex) autocorrelation function, normalized to $C_s(0) = 1$, is

$$C_s(t) = \exp[-(2\pi F_2 t)^2/2] \exp[-j2\pi F_1 t]$$

(22)

$N_{pc}$ is a natural unit of signal energy in optical applications and is related to the wideband carrier-to-noise ratio (CNR) by $N_{pc} = M(CNR)$ where $M = F_s T$ is the number of complex data points within the time series drawn from the lidar return if sampled at the Nyquist frequency. The correlation time [the 'bandwidth' of the Gaussian amplitude function in (22)] is $T_c = 1/(2\pi F_2)$.

In two limiting cases analytic expressions can be obtained for the standard deviation of return power estimates. In the first limit, in which spectral filtering is used to isolate the return signal, $F_2$ is much less than $F_s$, which is the usual mode of operation for lidars measuring Doppler wind shifts. If it is assumed that the Gaussian spectrum of (1) extends out to $3F_2$ then this condition might be quantified by writing $3F_2 < F_s/2$, or $f_2 = F_2 / F_s < 0.167$. If in addition the data time series is sufficiently long not to truncate the autocorrelation function, i.e., $3/(2\pi F_2) < T$, or $f_2 > 3/(2\pi M)$, then the approximation of Levin (1965) can be used to obtain an expression for the Cramer-Rao lower bound on the precision of the return power estimate. This result has been summarized by Arcese and Trombini (1971) whose formula can be written in the present notation as

$$f_{sd}(P) = \frac{\sigma(P)}{P} = \sqrt{\frac{2\pi}{m}} \left[ \int_{-\infty}^{+\infty} \frac{dx}{1 + (\sqrt{2\pi \alpha e^{-x^2/2}})^{-1}} \right]^{-1/2}$$

(23)

The bound is achieved if the total return signal energy integrated if necessary over many lidar pulses $N_{pc}$ is greater than about 100. The algorithm makes use of a maximum likelihood estimator that performs a matched filter operation on the spectrum of the return, the frequency at which the spectral peak occurs after filtering being the estimate of the Doppler shift. The return power is related to the height of this peak. The function used in the matched filter depends on $F_2$ so it is necessary to know the return bandwidth, which is determined in part by the bandwidth of the transmitted pulse and in part by the bandwidth of the Doppler shifts from scattering particles within the range gate; the former is known and the latter may not be, so a further condition is perhaps that the pulse bandwidth should be significantly greater than that of the atmospheric Doppler shifts. We assume that this is satisfied, and in addition that the pulse bandwidth is determined by the pulse duration, rather than, e.g., by chirp. Then the pulse width $T_{FWHM}$ defined as the full width of the power profile of the transmitted pulse at half height, for the Gaussian pulse with spectral profile given by (21), is $T_{FWHM} = (2\ln 2) / (2\pi F_2)$. Some achievable values for these parameters are given in Table 2. The Doppler search bands are typical of Doppler lidars, the range gate is chosen to be 100 m, and values for $f_2$ are chosen to satisfy the conditions just given.
Table 2: Typical values used for narrowband signal calculations: the range gate assumed is 100 m and the Doppler search band ±25 m/s

<table>
<thead>
<tr>
<th>wavelength</th>
<th>10 μm</th>
<th>2 μm</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_s$</td>
<td>10 MHz</td>
<td>50 MHz</td>
</tr>
<tr>
<td>$M = F_s T$</td>
<td>7</td>
<td>34</td>
</tr>
<tr>
<td>$(f_2)_{\text{min}} = 3/(2\pi M)$</td>
<td>0.068</td>
<td>0.014</td>
</tr>
<tr>
<td>$f_2$</td>
<td>0.1</td>
<td>0.05</td>
</tr>
<tr>
<td>$T_{\text{FWHM}}$</td>
<td>190 ns</td>
<td>75 ns</td>
</tr>
</tbody>
</table>

In the second limiting case, the signal bandwidth is made greater than the Doppler search band, $F_2 > F_s$, so that it is unnecessary to search for or optimize the detected signal by spectral filtering. Return power is then measured directly from the return time series by squaring the signal amplitude. The receiver bandwidth is set by $F_2$ and because this is large the range resolution is also optimized. To avoid truncation of the Gaussian signal spectrum in the absence of Doppler shift we should make the receiver bandwidth $B_R \geq 2(3F_2) = 6F_2$, so that if the Doppler search band is ±$F_d/2$ we need $B_R \geq 6F_2 + F_s$. We can then use Eq. (19) directly. If a series of $M_\nu$ independent observations are made within the range gate, then

$$\text{fsd}(P) = \frac{1}{\sqrt{M_\nu}} \left( 1 + \frac{1}{\text{CNR}} \right)$$  \hspace{1cm} (24)

The observations are independent if both the noise fluctuations and the signal fluctuations (speckle) are independent. Noise fluctuations are independent if the sampling period is a multiple of the reciprocal Nyquist frequency $1/B_R$. Signal fluctuations are independent if they are separated by more than the correlation interval. For the Gaussian signal, the correlation function after squaring is given by the square of $C_s(t)$ [Eq. (22)], so that the correlation time is given by $T_c/\sqrt{2}$. The number of independent values of the signal that can be obtained within the range gate $T$ is therefore $M_\nu = \sqrt{2T}/(3T_c)$.

For vertical sounding the Doppler search band can be reduced to, say, $F_s = \pm 5 \text{ m/s}$ so that the $B_R$ can be retained at the same value of 50 m/s used in the narrowband examples (where it equalled $F_s$) provided the pulse bandwidth $F_2$ is made less than $(B_R - F_s)/6 = 7 \text{ m/s}$. For the values given as examples in Table 2, $F_2$ has also been set to 5 m/s.
Table 3: Typical values used for broadband signal calculations: the range gate assumed is 100 m, the maximum radial wind speed is \( v_{\text{max}} = \pm 5 \) m/s, and the receiver bandwidth is \( \pm 25 \) m/s

<table>
<thead>
<tr>
<th>\text{wavelength, } \lambda \</th>
<th>\ 10 \mu m \</th>
<th>\ 2 \mu m</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_2 = (2/\lambda)v_{\text{max}} )</td>
<td>1 MHz</td>
<td>5 MHz</td>
</tr>
<tr>
<td>( T_{\text{FWHM}} )</td>
<td>190 ns</td>
<td>37 ns</td>
</tr>
<tr>
<td>( (T_s)_{\text{min}} = 3T_c/\sqrt{2} )</td>
<td>340 ns</td>
<td>68 ns</td>
</tr>
<tr>
<td>( T_s )</td>
<td>400 ns</td>
<td>80 ns</td>
</tr>
<tr>
<td>( M_s )</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

Some improvement can be obtained by sampling at the maximum rate, \( F_s \), in which case standard sampling theory (Davenport and Root, 1958) indicates that (4) can still be used if \( M_s \) is replaced by

\[
(M_s)_{\text{eff}} = \frac{M}{1 + 2 \sum_{k=1}^{M-1} \left( 1 - \frac{k}{M} \right) |C_S|^2 \left( \frac{k}{F_s} \right)}
\]  

(25)

For the 10 \( \mu m \) system this leads to \( (M_s)_{\text{eff}} \approx 3.8 \) and for the 2 \( \mu m \) system to \( (M_s)_{\text{eff}} \approx 8.8 \).

Values of the fractional standard deviation of return power estimates are plotted as a function of \( N_{pc} \) calculated using Eqs. (23) - (25) and the data of Tables 2 and 3, are given in Figs. A.5. The curves are similar in that at low signal \( \text{fsd}(P) \) increase rapidly (as \( 1/P_r \) not \( 1/\sqrt{P_r} \)) and at high signal it saturates, so that from a single range gate \( \text{fsd}(P) \) is not less than about 0.3. The former is the reason why heterodyne power measurements have a threshold (i.e., the CNR should usually be at least about 1) and the latter the reason why multiple return pulses must be used. It might be noted that the pulse durations to be used, given as \( T_{\text{FWHM}} \), are for the 10 \( \mu m \) system very similar in Tables 2 and 3, so that in effect we are comparing results using the filtering and sampling approach as applied to the same system. In this case use of maximum rate sampling [Eq. (25)] is advantageous and leads to results that are not very different from those obtained using filtering. For the 2 \( \mu m \) system on the other hand there is very little difference between \( M_s \) and \( (M_s)_{\text{eff}} \) and especially at low CNR use of filtering could generate a useful enhancement of performance. The ratio of \( \text{fsd}(P) \) obtained from independent sampling to that using filtering is also shown in Fig. A.5.

In the simulations used in this report the parameters assumed are those of Table 3. For the 10 \( \mu m \) heterodyne systems, the independent sample expressions for \( \text{fsd}(P) \) are used. From Fig. A.5(a), this is only slightly conservative. For the 2 \( \mu m \) systems the improvement to be expected from filtering at low CNR [Fig. A.5(b)] is taken into account. In practical application the independent sample approach is clearly the simpler - it leads to the lowest data rates and processing is straightforward.
Figure A.5(a): Fractional standard deviation of power measurements for single rangegate in return from CO$_2$ lidar with parameters of Tables 2 and 3. Estimates obtained from (i) filtered ML algorithm (heavy solid line), (ii) independent sampling (heavy dashed line) and (iii) full sampling using effective number of independent data points (open squares). The ratio of the (iii) to (i), which shows the advantage to be gained from Doppler processing) is also shown as the dotted line.

Figure A.5(b): As Fig. A.5(a) except for 2 µm lidar.
5.6 Column content and range-resolved concentrations

It can be seen from the results of the previous subsection and in particular those of Figs. A.5, in which $\text{fsd}(P)$ from a single pulse and a single range-gate is never less than about 0.3, that the need to improve the precision of return power estimates is intrinsic to heterodyning. In direct detection DIAL, the SNR and therefore $\text{fsd}(P)$ is determined only by the power of the return [Eq. (20)] which may be adequate if a high transmitted pulse power is used and at short range, but usually some improvement is required. In both cases data has to be combined from surrounding measurements by using the return from separate transmitted pulses and/or from neighboring ranges in the return from each pulse. This is a two dimensional range/time processing problem but we shall separate it into two parts, first processing returns obtained from a given range-gate over time which leads to provisional column content estimates, and then processing in range which leads to smoothed column content and range-resolved estimates.

The usual way to approach such estimation is through averaging. If the $j$th measurement $y(j)$ is described by the measurement equation

$$y(j) = x(j) + v(j)$$

where $x$ is the unknown and $v(j)$ is a zero mean noise term, then an estimate $x^*$ of $x$ is obtained by averaging using

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{N} y(j) = x^* + \bar{v}$$

so that for large $N$ where $\bar{v} \to 0$, $x^* \to \bar{y}$. This approach is unsuitable for use where $x$ is not constant. If $x = x(j)$ and is not constant over the period of measurement, which is often the case for atmospheric parameters, then its average may be difficult to interpret. Ideally we may want to follow the fluctuations of $x(j)$. Variation in the parameter being estimated is even more of a problem if the measurement is a nonlinear function of the parameter. For example, if the return power is the measured parameter $y$, and if it is the column concentration that is the unknown parameter $x$, then the measurement equation for (e.g.) direct detection has the form

$$y(j) = e^{x(j)} + v(j)$$

(assuming other lidar equation parameters do not vary in time and omitting constants); averaging the $y(j)$ will give the average of $\exp[x(j)]$, not the exponential of the average of $x(j)$. It is necessary to know the statistics of $x(j)$ in order to derive the one from the other.

In order to take account of time dependence the data can be treated recursively, i.e., point by point as it comes in, rather than as a block. The average after $j$ data points have been collected can for example be written

$$\bar{y}(j) = \bar{y}(j-1) + \frac{1}{j}[y(j) - \bar{y}(j-1)]$$

(29)
This enables the average of \( y(j) \) to be determined in terms of the average after \( j-1 \) data points and the new measurement \( y(j) \). It can be used for any \( j \) and is easily derived from Eq. (27) by induction. In principle, if \( y \) is linear in \( x \) and it is believed that \( x \) only varies slowly then the first approach might be to form a moving average of not the whole data set but only of small subsets. This however would fail to take advantage of the recursive approach to data processing (a different block of \( j \) data points has to be processed for each update) and it is easier to form the exponentially weighted moving average using

\[
\overline{y(j)} = \frac{1}{k} [y(j) - \overline{y(j-1)}] + \overline{y(j-1)}
\]

(30)

where \( k < j \). \( k \) here is constant and gives the 'memory' of the filter (the number of data points effectively averaged over). \( 1/k \) is the 'gain' of the filter (the relative weighting of the new measurement).

The next problem is to fix a value for \( k \), which is best not set \textit{a priori} but should depend on the rapidity of the rate of change in \( x \). The first step is to introduce a model for the time dependence of \( x \). If \( x \) is constant it can be described in recursive terms by

\[
x(j) = x(j-1)
\]

(31)

for all \( j \). If there is no knowledge of the actual time dependence then the simplest generalization to the constant signal model is

\[
x(j) = x(j-1) + w(j)
\]

(32)

where \( w(j) \) is a white zero mean noise term. Eq. (32) describes a random walk, and is the simplest example of a stochastic model. Stochastic models are an alternative to various deterministic models that might be fitted to the data, \textit{e.g.}, high-order polynomials, or cubics fitted using a splining procedure. An advantage of stochastic modeling here is that it is both commonly used in time series estimation and that it is closely related to the modeling commonly used in turbulent atmospheres. For example, the random walk of Eq. (32) has the spectrum \( 1/f^2 \). Such a spectrum has been observed by Melfi (1994) in the concentration of water vapor measured using Raman scattering and can obviously serve as a 'quadratic' approximation to the Kolmogorov \( 1/f^3 \) spectrum predicted for passive additives in turbulent atmospheres. It is in general difficult or impossible to model fractal spectra like Kolmogorov's using time series expressions like Eq. (32).

Given the system model (32), we want to combine with the measurement model (26) in such a way that the updated estimate is given by a simple recursion similar to (30), \textit{i.e.},

\[
x^*(j) = x^*(j-1) + K(j) [y(j) - x^*(j-1)]
\]

\[
= K(j) y(j) + [1 - K(j)] x^*(j-1)
\]

(33)

The second form of this equation shows that \( y(j) \) and the former estimate \( x^*(j-1) \) are being combined linearly. It is well-known that an optimal way of doing this, one that minimizes the variance of the updated estimate, is to weight \( y(j) \) and \( x^*(j-1) \) by their inverse variances. If we write for the measurement noise \( \text{var}[y(j)] = \text{var}[v(j)] = R(j) \) and if we assume that immediately before the measurement we know \( \text{var}[x(j-1)] = \Lambda^0(j) \), then this weighting leads to
is the updated variance of the estimate $x^*(j)$ immediately after the measurement. Immediately before the next measurement this variance has increased because according to the system model (32) there has been an increase in the uncertainty in $x^*(j)$. The variance becomes

$$\Lambda^{(-)}(j+1) = \Lambda^{(+)}(j) + Q(j)$$

where $Q(j) = \text{var}[w(j)]$. The equations (33) to (35) are all that is needed to process the data given the assumptions about the models contained in Eqs. (26) and (32). If the models do not fit the data, then estimates will still be obtained but they will be sub-optimal.

These equations are applied to each measurement in turn to form a series of estimates of $x(j)$. The approach for which we have given this example is equivalent to Kalman filtering and $K(j)$ is known as the Kalman gain (Gelb, 1974). It can be used for time series that are nonstationary - we already have a random walk model for $x(j)$ that is nonstationary and in lidar applications the measurement model is likely to be, for $R(j)$ may vary with $j$. This is useful when dealing with range-smoothing for example, because the noise increases with range. The only uncertainty is whether the system model Eq. (32) is adequate to describe $x(j)$ and the only unknown is $Q(j)$. There are standard self-consistency tests that can be performed on the output to determine whether the system model is adequate and there are a number of techniques for estimating $Q(j)$. The $Q(j)$ characterize the variability of the data and in part determine $K(j)$, so that their estimation on-line has the effect of making the filter adapt itself to the data.

The approach can be generalized to a whole range of estimation problems and there is a large literature to draw on. For the present purpose two particular problems are of interest (Rye and Hardesty, 1989). Because DIAL makes use of the difference between returns at (at least) two wavelengths, we are concerned with situations in which there is more than one unknown, in which case the unknowns are collected into a vector and the equations become matrix manipulations. We can also note a straightforward first order approximation for handling nonlinear measurement equations like (28). This implies that the measurement $y(j)$ has to be combined not with $x^*(j-1)$ but with $\exp[x^*(j-1)]$. If $\text{var}[x^*(j)] = \Lambda^{(-)}(j)$ then $\text{var}[\exp(x^*)] \approx (\frac{de}{dx})^2 \text{var}[x^*(j)] e^{2x^*} \text{var}[x^*(j)]$, and substituting this in (34) yields

$$K(j) = \frac{e^{2x^*(j-1)} \text{var}[x^*(j-1)]}{R(j) + e^{2x^*(j-1)} \text{var}[x^*(j-1)]}$$

Moreover, for (33) we can likewise linearize around the estimates, using $\exp[x^*(j)] - \exp[x^*(j-1)] \approx (\frac{de}{dx})[x^*(j) - x^*(j-1)] = e^x(x^*(j) - x^*(j-1))$ to obtain
which is the update equation for the estimate of what is known as the extended Kalman filter. Eq. 
(37) could be used to determine the column absorption from measurements of the return power at 
a single range gate.

This procedure has not been used in the simulations reported here. Instead the simplifying 
assumption has been made that, over a large number of pulses transmitted at two wavelengths within 
a short time interval, reliable estimates of the logpower ratio or column content absorption needed 
for DIAL can be formed by averaging. It is believed that the problems associated with lidar equation 
parameter fluctuation will not prevent such estimates being obtained over short periods of time, i.e., 
about a few seconds. These estimates are not however good enough to give range-resolved 
concentrations by simple differencing (see the examples in Figs. 5 and 6 discussed in section 2) and 
a range smoothing algorithm must be used. The principles underlying the range-smoothing algorithm 
used here, which is based on the Kalman filter approach, are as follows. The column content at range 
increment \( j \) is \( x_1(j) \). The range-resolved concentration between increments \( j \) and \( j-1 \) is 
\( x_{r-1} = x_1(j) - x_1(j-1) \) and is itself modeled as a random walk, so that the system model becomes

\[
\begin{align*}
  x^*(j) &= x^*(j-1) + \frac{e^{x^*(j-1)}}{R(j) + e^{2x^*(j-1)}} \frac{\text{var}[x^*(j-1)]}{\text{var}[x^*(j-1)]} \\
  & \quad \times [y(j) - e^{x^*(j-1)}]
\end{align*}
\]  

(37)

The filter equations are applied assuming knowledge of the variance of the column content estimates 
obtained by averaging. These column content estimates are regarded at this stage as the 
measurements, and are described by \( y(j) = x_1(j) + \nu(j) \). The variance of the range-resolved 
concentration model, \( \nu(j) \), is arbitrarily assumed to be proportional to the square of the water vapor 
concentration in the US standard atmosphere model (see following section), the constant of 
proportionality being determined by maximizing the likelihood of the series of estimates. This 
assumption does not seem to be critical, but is consistent with our expectation that as the water 
concentration becomes lower it also becomes less variable. Filter outputs always reflect past inputs 
and if (for example) the processing starts with data from short range and proceeds to longer ranges 
then the range-resolved concentrations obtained in this way are biased by the measurements made at 
lower ranges. This is overcome by use of smoothing, i.e., constructing a second filter that works in 
the other direction, starting from long range. The outputs of the two filters are then combined 
optimally using weighting determined by the inverse of their variance estimates.

6 Appendix 2: Details of the simulations

6.1 Geometry

The lidar is assumed to be ground-based and operated from a platform of initial height \( h_0 \) above 
the ground. The altitude of an observation at range \( r \) is then
\[ h = h_0 + r \sin \theta \]  

(39)

For the simulations used here the elevation angle \( \theta \) is always 90°. The ranges at which calculations are made correspond to the data sampling points, \( i.e., \ r = (i + 1) \Delta r \) where \( i \) is an integer (from 0) and \( \Delta r = c/(2F_s) \), \( F_s \) being the sampling frequency.

6.2 Backscatter coefficient

The volume backscatter coefficient \( \beta \) is modeled using the model of Kavaya et al., 1989, who fitted various empirical values obtained from CO\(_2\) lidar observations and assumed that \( \beta \) varied as \( \lambda^2 \) to extrapolate these to 2 \( \mu \)m wavelengths. The model may consequently be somewhat more realistic at 10 \( \mu \)m and slightly optimistic for 2 \( \mu \)m, although of course actual values vary widely and any model is only a guide. The formula used is

\[
\beta_{(m \ sr)^{-1}} = \left( \frac{2.1}{\lambda_{\mu m}} \right) \times 10^{3.52 \exp[-h_{\text{meter}} / 4967]} - 9.46
\]

(40)

where the subscripts on \( \beta \), \( \lambda \), and \( h \) indicate units, and values for \( \beta \) are plotted in Fig. A.6.

6.3 Refractive turbulence

Refractive turbulence is modeled for heterodyne lidar using the (conservative) model based on bistatic geometry. The coherence area is given by

\[
\rho = \left[ 1.45 \left( \frac{2\pi}{\lambda} \right)^2 \int_0^r C_n^2(x) \left( 1 - \frac{x}{r} \right)^{5/3} \, dx \right]^{-0.6}
\]

(Yura, 1979) where the variation of the structure parameter with altitude is modeled using

\[
C_n(h_{\text{meter}})^{-2.3} = 10^{-12} \ h_{\text{meter}}^{-4/3}
\]

(42)

A ground-based lidar with vertical-pointing beam is most affected by the turbulence near the ground (see section 5.3.3), which can also in practice change by orders of magnitude from one time to another. Again therefore these values should be regarded as only a guide to performance.

6.4 Physical parameters and species concentrations

The variation with altitude of pressure, temperature, density, and the concentrations of the only species in the ambient atmosphere that have significant absorption at these wavelengths, water vapor, CO\(_2\), and O\(_3\), were taken from the U.S. standard atmosphere model. This gives values at 1 km altitude intervals, and these were fitted using a cubic spline routine to obtain values at intermediate levels. Resulting profiles are shown in Figs. (A.7) and (A.8).
Figure A.6: Backscatter coefficients as function of altitude for 10μm (heavy solid line) and 2 μm (solid line)

Figure A.7: US standard atmosphere values for altitude dependence of pressure (heavy solid line) and temperature (solid)

Figure A.8: US standard atmosphere values for altitude dependence of concentrations of water vapor (heavy solid line), CO₂ (heavy dashed line) and ozone (solid)
In order to make estimation of water vapor concentrations more challenging, the water vapor profile was randomized about the mean values given by the standard atmosphere. The fluctuating component of the water vapor concentration was described as a discrete exponentially-correlated variable. That is, the concentration model at the \( i \)th range increment was

\[
n(i) = \overline{n(i)} + \Delta n(i) \tag{43}
\]

where \( \overline{n(i)} \) is the mean value given by the standard atmosphere, and the fluctuating component was given by

\[
\Delta n(i) = \Delta n(i-1) \ e^{-\frac{[r(i) - r(i-1)]}{r_{H2O}}} + w(i) \tag{44}
\]

Here \( r_{H2O} \) is the distance along the lidar line-of-sight characterizing the length scale of the fluctuations and \( w(i) \) is a white noise term of variance

\[
\text{var}[w(i)] = \text{var}[n(i)] \sqrt{1 - e^{-\frac{[r(i) - r(i-1)]}{r_{H2O}}}}
\]

The variance \( \text{var}[n(i)] \) characterizes the level of the fluctuations. Examples of the water vapor concentrations obtained using these formulas are given in the text. The spatial spectrum of the fluctuations is Lorentzian, which approximates a \( 1/f^2 \) spectrum (like the random walk) over intervals smaller than \( r_{H2O} \).

### 6.5 Line absorption

The attenuation \( T(j,v) \) at wavenumber \( v_w = 1/\lambda \) for returns from range increment \( r(j) \) can generally be divided into terms arising from particle scattering and molecular absorption, and in the infrared the latter predominates. For the lidar round-trip we write

\[
\alpha_k(i,v_w) = \eta_k(i) \sum_m \sigma(\omega_m,v_w)
\]

where the \( \omega_m \) are the wavenumbers of absorption lines. Assuming only pressure broadening, the cross-sections \( \sigma_k(\omega_m,v_w) \) have the Lorentzian profile

\[
T(j,v_w) = \exp\left[-2\sum_{i=0}^{i=j} \alpha(i,v_w)\Delta r(i)\right] = \exp\left[-2\Delta r \sum_{i=0}^{i=j} \alpha(i,v_w)\right] \tag{47}
\]

(the final expression if as here all range increments are equal). The absorption coefficient \( \alpha(i,v_w) \) at the \( i \)th increment is the sum of coefficients for each absorbing species, \( \alpha(i,v_w) = \sum \alpha_k(i,v_w) \). For the \( k \)th species of concentration (in particles per unit volume) \( \eta_k(i) \),

\[
\sigma(\omega_m,v_w) = \frac{S(T) \ \gamma(p,T)}{\pi [\gamma(p,T)^2 + (\omega_m - v_w)^2]} \tag{48}
\]

where the pressure \( p \) and temperature \( T \) dependence of the linewidth \( \gamma \) and line strength \( S \) are given by
\[
\gamma(p,T) = \gamma_m \left( \frac{P_k(i)}{P_0} \right) \left( \frac{T_0}{T(i)} \right)^{n_m} \\
S(T) = S_m \left[ \frac{T_0}{T} \right]^j e^{-\frac{h_c E_0}{k} \left( \frac{1}{T(i)} - \frac{1}{T_0} \right)}
\]

(49)

The \( \gamma_m, S_m, E_m, n_m, \) and \( \omega_m \) for each line of each species are listed in the well-known HITRAN data tapes prepared at the former Air Force Geophysics Laboratory, now the Geophysics Division of the Phillips Laboratory (e.g., Rothman et al., 1987). The coefficients \( J \) have the values 1.0 (water vapor), 1.5 (CO\(_2\)), and 1.0 (O\(_2\)). The standard constants \( P_0 = 1013 \) mbar and \( T_0 = 296 \) K.

6.6 Water vapor continuum absorption at 10 \( \mu \)m

In the mid-infrared there is a continuum absorption arising from water vapor which has the effect of reducing atmospheric transmission at these wavelengths and has a quite different temperature dependence from the absorption lines. These are described by the numerical equations of Selby et al., 1976, developed for the LOWTRAN program:

\[
\alpha(i, \nu_w) = c_s(\nu_w T_0) W(i)
\]

\[
W(i) = \frac{w(i)}{10} \frac{P_{H2O}(i)}{P_0} \left[ 6.08 \left( \frac{T_0}{T(i)} - 1 \right) + 0.002 \left( \frac{P(i)}{P_{H2O}(i)} - 1 \right) \right]
\]

(50)

where the wavenumber \( \nu_w \) is in cm\(^{-1}\), and the density \( w(i) \) is in gm/m\(^3\). \( P_{H2O}(i) \) is the partial pressure of water vapor at the \( i \)th range increment.

6.7 Examples of absorption spectra

Figs. A.9 and A.10 show examples of ambient atmosphere absorption spectra at some wavelengths of interest for infrared DIAL, calculated using the formulas of the two previous sections. The conditions assumed when constructing these plots were uniform temperature of 290 K, sea-level pressure (1013 mbar), and 50% relative humidity (corresponding to a water vapor density of about 7 gm/m\(^3\)). When determining the overall absorption at any wavelength, it was found necessary to include contributions from spectral lines up to 10 cm\(^{-1}\) distant in order to take account of the overlap of the Lorentzian lines, the wings of which are weak but highly extended.

At around 10 \( \mu \)m, the absorption has been calculated at the discrete transmission wavelengths of the CO\(_2\) laser using the common isotope C\(^{12}\)O\(_{16}\). The results are represented in the bar chart of Fig. A.9. The two principal components of the absorption arise from the water vapor continuum and from CO\(_2\) itself. The latter is due to absorption by the same transitions that produce the laser emission and
Figure A.9: Absorption spectrum for CO$_2$ laser lines, showing, from the bottom, water vapor continuum (shaded, at base), CO$_2$ lines (open), ozone band (dotted) and water vapor lines (filled). The absorption in dB/km of those water lines that extend above the graph boundary is shown at the top of the graph.

Figure A.10: Absorption spectrum for 2 µm Th$_3$Lu$_2$YAG laser showing line spectra of water vapor (heavy solid line) and CO$_2$ (solid line).
would not be present if another isotope of CO\textsubscript{2} was to be used in the laser. These contributions give rise to attenuation of between about 0.2 and 0.8 dB/km. Above these components that are present at each wavelength, there is some weak ozone absorption in the 9 \mu m transmission band, and a number of water lines across the spectrum of varying strength. The two frequencies used for the DIAL simulations are at 974.6 cm\textsuperscript{-1} (the 10R18 line) and its neighbor (to ensure about the same CO\textsubscript{2} absorption) at 975.9 cm\textsuperscript{-1} (10R20), the latter coinciding with a strong water line.

The 2 \mu m laser is continuously tunable and a small section of the absorption spectrum is shown in Fig. A.10 around the frequencies used in the simulations, which were 4943.9 cm\textsuperscript{-1} (2.023 \mu m) and 4948.1 cm\textsuperscript{-1} (2.021 cm\textsuperscript{-1}), the latter again coinciding with the water absorption. These wavelengths are accessible to the Tm:Lu\textsubscript{2}YAG laser. There is in this spectral region a dense line structure arising from both water vapor and CO\textsubscript{2}, resulting in a background absorption of about 0.2 dB/km and providing the opportunity for monitoring either compound using either weak or strong lines.

6.8 Summary of lidar parameters

The principal differences between lidar parameters at the two wavelengths are that the 2 \mu m systems are more susceptible to degradation from refractive turbulence (although this is not likely to be a large problem for vertical line of sight), aerosol backscatter is greater at 2 \mu m than at 10 \mu m (this is compensated by a lower level of LO shot noise for the latter), the greater receiver bandwidth of 2 \mu m systems which increases the receiver noise but also enables a higher degree of averaging of speckle noise, and absorption of the broadband water vapor continuum which augments the line absorption from molecules at 10 \mu m.

7 Conclusions

Our results, as exemplified in the figures here, suggest that the reluctance to use infrared DIAL based on the likely performance of direct detection 10 \mu m lidars is well-founded but that compact heterodyne systems operating at either 2 \mu m or 10 \mu m with current technologies could provide useful, eye-safe, day and night measurements at altitudes up to about 3 km. The statistical uncertainty of the water vapor estimates obtained is predicted to be usually less than about 20\% or 1 gm/m\textsuperscript{3}, the smoothing distance less than about 200 m, and the time resolution 10 secs. The 2 \mu m lidars that could be used for this purpose would have optics diameters as small as about 5\". At higher altitudes, up to 5 km, it is necessary to rely on long averaging times and it would have to be demonstrated whether this is practical. As higher power laser technology develops, routine profiles through most of the troposphere will be feasible.

Development of the 2 \mu m lidar technologies considered here is presently proceeding for Doppler applications. There are at present no obvious show-stoppers preventing ultimate extension of this technology to turn-key water vapor DIAL. Practical development of this additional capability and testing of this technique would require only moderate modification of the existing NOAA system.
With partial funding from ARM, development of the high PRF MOPA CO\textsubscript{2} technology is currently proceeding at ETL for application to multi-wavelength remote sensing of cloud microphysics. With modest modification, this system could be made suitable for a testbed for coherent DIAL measurement of water vapor and other species. Again, there appears to be no obvious showstoppers preventing ultimate extension of this technology to turn-key water vapor DIAL.

In the course of this study we have made preliminary calculations which suggest the DIAL technique may also be useful for measuring temperature profiles. This is useful both to further improve the accuracy of the DIAL water vapor concentration retrievals, and possibly, as an additional method of providing continuous eye-safe, day and night temperature measurements suitable for thermodynamic profiles.

References for Part 1


Cummins, H.Z., and R.L Swinney, 1970: Light-beating spectroscopy, Prog. in Optics, 8, 123.


Evaluation of tropospheric water vapor profiling using eye-safe, infrared differential absorption lidar

Part II: Diode-based lidars

8 Introduction

A thorough understanding of tropospheric water vapor is essential for forecasting weather and modeling radiative transfer. This project has been to examine methods based on differential absorption lidar (DIAL) for obtaining high quality vertical profiles of water vapor with high spatial and temporal resolution and complete diurnal coverage, with the eventual goal of instrumenting the CART sites. We focus on the development of a compact, automated, eye-safe, ground-based and inexpensive differential absorption lidar (DIAL) to profile atmospheric water vapor. Such a system, requiring little adjustment in the field, could be widely deployed to supplement twice-daily radiosonde profiles or used in arrays to measure water vapor transport and horizontal water vapor variability with a continuous data stream from a known location. The lidar will be based on an all solid-state laser with high spectral resolution, and good frequency and amplitude stability. Instrument features will include daytime measurement capabilities, a detection range of 3-6 km, a spatial resolution of 100 m, a temporal resolution of 15 minutes, and an error of less than 1 g/kg.

In the Part I of this report we examined the theoretical feasibility of performing water vapor DIAL measurements with existing lidar systems using either direct or heterodyne detection at 2 and 10 µm. Recent advances in diode laser technology suggest that a simpler and more compact instrument might be designed at shorter wavelengths. This second part of the report discusses the results of new lidar simulations between 0.8 and 1.5 µm and the technology options for a low-cost, diode-based water vapor lidar in the near infrared. Specifically, the proposal extension identified the following tasks:

1. Develop realistic models to predict performance of laser diode water vapor profiling systems, including realistic approximations for system power, detector count rates, atmospheric aerosol backscatter, and water vapor variability.

2. Assess current technology and predict potential future technological advances that support laser diode water vapor measurements, examining such issues as Geiger mode APDs versus photomultiplier tubes, properties of diode sources, and utility of MOPA configurations for added energies.

3. Assess the relative advantages and disadvantages of pulse compression (pseudo-random code modulation) versus single pulse systems.

4. Perform laboratory experiments to assess diode laser stability, tunability, and spectral characteristics with respect to DIAL water vapor measurements.
5. Prepare a final report documenting the investigation and provide an assessment of the feasibility of diode laser-based systems for water vapor studies. Compare the relative advantages and disadvantages of diode systems with the coherent systems studied under the previous proposal.

The results of these studies are discussed in the following sections.

9 Simulation

A computer simulation was written to model DIAL systems for three types of receivers: pulsed direct detection, pseudo-random (rm) code modulation, and heterodyne detection. The performances of these systems are examined in the near-infrared, between 0.8 and 1.5 μ while parameters such as backscatter, absorption, and wavelength are varied.

9.1 Parameters

The basic lidar equation which defines the received power, \( P_r(i) \), from altitude step \( I \) is generally written as follows:

\[
P_r(i) = P_0 \frac{c}{2} \tau_{\text{pulse}} \pi \frac{\rho_{\text{back}}}{R^2} \beta(i) \beta_{\text{ext}}(i) \eta \int_0^R f_{\text{bw}} e^{-2 \int_0^R (\gamma + \alpha) dr}.
\]

Here, \( \alpha \) and \( \gamma \) are the absorption coefficients for the water and non-water components of the atmosphere, and \( \beta_{\text{ext}}(i) \) is the extinction due to backscatter. Definitions and typical values of the other parameters are listed in Table 2.

For the purposes of this simulation, the lidar equation is rewritten as:

\[
P_r(i) = \left[ \frac{P_0}{P_{\text{off}} - P_{\text{on}}} \right] \left[ \eta \int_0^R \beta_{\text{ext}}(i) \right] \left[ \int_0^R (\gamma + \alpha) dr \right] \left[ \pi \theta_{\text{range}}^2 \right] \beta(i).
\]

In this equation, the first factor is the average power (single pulse energy / time of one range step), the second term contains the attenuation factors, and the final factor is the backscatter into the receiver. Here \( \pi \theta_{\text{range}}^2 \) is the solid angle accepted by the receiver. The water absorption coefficient is calculated from \( \alpha(i) = -\ln \left[ \frac{P_0(i)}{P_0(i) - P_{\text{off}}(i)} \right] \), where \( P_0(i), P_{\text{on}}(i), \) and \( P_{\text{off}}(i) \) are the on-line, off-line and background return signals from range \( I \).

The time scales are chosen as follows. The vertical resolution of 100 m corresponds to time steps of 0.667 μs (2*100 m = c*0.667 s). The maximum desirable pulse repetition frequency is about 6 kHz (166-μs period) which permits an unambiguous measurement range of 25 km (2*25 km = c / 6 kHz). For direct detection, the use of lower frequencies will increase the signal-to-noise ratio (SNR) by both increasing the pulse energies and decreasing the amount of background integrated per pulse. The "dead time" which is present after the pulse has covered the unambiguous range, but before the next pulse, can be used for background measurements.
For the rm-code case the 6-kHz frequency is slightly modified because there must be $2^n$ bits per cycle. A code of 256 pulses, each corresponding to a vertical resolution of 100 m, results in a cycle repetition frequency of 5.859 kHz. Since there is no "dead-time" in the rm-code case, it would not be an advantage to lower the cycle frequency; this would just extend the unambiguous measurement range.

The optics are chosen so that the transmitted beam is nearly collimated, a reasonable assumption in the near field because amplified diode laser beams can be near the diffraction limit. [Mehuys, Goldberg and Welch] For direct detection this limits the required acceptance angle for the receiver and reduces the collected sky radiation. [Halldörsson and Langerholc] The return signal is coaxial with the transmitted beam and the acceptance angle is chosen so as to aperture the signal from the closer ranges, thereby decreasing the required dynamic range of the detector. In the heterodyne case, one speckle can be detected at most, and the acceptance angle is the diffraction-limited angle, 

$$\theta_1 = \frac{\lambda}{2r_{\text{beam}}}.$$ 

The received signal due to sky background is

$$P_{\text{sky}} = L_{\text{sky}} f_{\text{od}} f_{\text{bw}} \eta \pi r_{\text{beam}}^2 \theta_1^2,$$

with wavelength: $L_{\text{sky}}(\lambda) = Ae^{-2.2\lambda}$. 

The luminosity is assumed to vary exponentially

Table 2. Parameters for lidar simulation.

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>TYPICAL VALUE AT 828 nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>acceptance angle $\theta_0$</td>
<td>200 $\mu$rad</td>
</tr>
<tr>
<td>backscatter (at 0 km) $\beta(0)$</td>
<td>$10^{-6}$ m$^{-1}$ sr$^{-1}$</td>
</tr>
<tr>
<td>detector efficiency $\eta$</td>
<td>0.3</td>
</tr>
<tr>
<td>detector dark count</td>
<td>20 cps</td>
</tr>
<tr>
<td>integration time</td>
<td>15 min</td>
</tr>
<tr>
<td>laser beam radius $r_{\text{beam}}$</td>
<td>0.18 m</td>
</tr>
<tr>
<td>laser pulse duration $T_{\text{pulse}}$</td>
<td>20 ns</td>
</tr>
<tr>
<td>laser pulse repetition rate $\text{prf}$</td>
<td>6 kHz</td>
</tr>
<tr>
<td>maximum range $R$</td>
<td>6 km</td>
</tr>
<tr>
<td>optical filter bandwidth $f_{\text{bw}}$</td>
<td>0.1 nm</td>
</tr>
<tr>
<td>optical filter transmission $f_{\text{od}}$</td>
<td>0.3</td>
</tr>
<tr>
<td>sky background radiance $L_{\text{sky}}$</td>
<td>0.01 W/(sr m$^2$ nm)</td>
</tr>
<tr>
<td>transmitted average power $P_o$</td>
<td>350 mW</td>
</tr>
<tr>
<td>vertical resolution $r_{\text{stop}}$</td>
<td>100 m</td>
</tr>
</tbody>
</table>

Atmospheric absorption spectra as a function of altitude are obtained from HITRAN-PC routines.
Figure 32. Maximum permissible exposure limits for a 17.8-cm radius beam. The cw values apply to the rm-code method.

Figure 33. Water vapor density profiles for AFGL tropical and U.S. standard model atmospheres.
assuming Voigt line shapes. There are separated water absorption lines of varying magnitude throughout most of the near infrared region which are suitable for the DIAL measurements. However, the absorption line strengths are too small for water vapor DIAL in the regions of 845 to 885 nm and 1.03 to 1.066 μ, and too strong between 932 and 960 nm. For the DIAL system, the selected absorption lines should exhibit not only the desired absorbances, but also minimal pressure and temperature broadening. Absorption profiles versus altitude are generated for both the AFGL tropical or US standard atmospheric models. (Fig. 33)

In this lidar simulation, the laser linewidths are neglected. As has been discussed by Liu and Reagan at the University of Arizona (contractors) and Cahen and Mengie, errors can occur in real DIAL measurements if linewidth is not take into account due to the decrease in pressure broadening with altitude.

The laser wavelengths should be in a region of limited continuum absorption because this absorption decreases the signal returns in both the on-line and the off-line beams and has a strong temperature dependence. In the near infrared, the continuum is due almost entirely to the wings of strong water vapor lines, and is strongest in the regions near 0.94, 1.12 and 1.36 μ.

The optimum absorption line strength is assumed to be that which permits the largest range of detection. The ideal line strength at 830 nm was assessed by comparing the results of the DIAL simulations for varying laser and water vapor line strengths and under different atmospheric conditions (AFGL tropical or US standard atmospheric models). Figure 34 shows the range where the SNR of the measured absorption coefficient crosses a threshold of 3 for a number of absorption lines as a function on vertical atmospheric transmission. The figure indicates that a wavelength selected to give ~80% transmission over a 1-km path (vertically) would provide near optimal performance. A water vapor line which has ~85% transmission through 1 km in the US standard atmosphere will still work well in the tropical atmosphere where it will have ~65% transmission.

In the model backscatter profile, the top of the boundary layer is near 2 km (Fig. 35). The backscatter profiles are based on a composite of previous measurements and models. The backscatter is assumed to vary as $\lambda^{-1}$. Temporal variability of the atmosphere is neglected. Assuming the Mie regime, the extinction due to scatter at range step $I$ is given by: $\beta_{ext}(\lambda) = \prod_{j=0}^{I-1} (1 - 40 \beta(\lambda)r_{step})^2$.

9.2 Bias and atmospheric variability

The effect of atmospheric fluctuations including refractive turbulence have not been taken into account in any detailed way in the current calculations. Even if, as assumed here, the ultimate goal is estimation of the average concentration of water vapor over a period of 15 minutes, short-term estimates are needed over a period comparable with that of fluctuations in the lidar equation parameters. The reason is that the measurement is nonlinear. Consequently, although the instantaneous return power is related exponentially to the instantaneous absorption in a DIAL measurement, it does not follow that the average return power over a long period is likewise related exponentially to the average absorption. Short-term estimates of the time-dependent absorption must be processed to obtain the average, or any other moment of the concentration that is required. A
Figure 34. Vertical range where SNR crosses a threshold of 3 as a function of transmissivity. This indicates that a wavelength selected to give ~85% transmission over a 1-km path length would provide near optimal performance.

Figure 35. Assumed backscatter cross-section at 830 nm.
Table 3. Signal levels from different ranges for the parameters in Table 2. Optical and detector losses are assumed and units are counts per range gate per pulse at the detector. The sky background is 0.34 counts per gate per pulse.

<table>
<thead>
<tr>
<th>Average beam power</th>
<th>Altitude (m)</th>
<th>On-line counts / gate / pulse</th>
<th>Off-line counts / gate / pulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>350 mW</td>
<td>150</td>
<td>75</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>3050</td>
<td>4.3</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>5950</td>
<td>0.16</td>
<td>0.99</td>
</tr>
<tr>
<td>10 mW</td>
<td>150</td>
<td>2.2</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>3050</td>
<td>0.12</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>5950</td>
<td>0.047</td>
<td>0.028</td>
</tr>
</tbody>
</table>

The simplest approach is to base the absorption estimate on short-term averages of the return power. Such estimates will necessarily be noisy, but it is more important that they should be unbiased.

The problem is whether the low level of signal that is available in these shorter periods is sufficient for an unbiased estimate to be obtained. Bias can arise from small sample effects in statistics; e.g., an estimate of a ratio may be biased if formed from noisy estimates of the numerator and the denominator. It also arises from the need to reject a fraction of these estimates (the rejection ratio); e.g., some short-term estimates of return power will be negative after subtracting the mean background and cannot be used to form a logarithm.

The pulsed direct detection systems considered here make use of photon counting and the primary sources of noise are shot noise in the signal and the sky background light. Suppose that the counts from a given range and over a certain interval at the absorbed and reference wavelengths are \( n_1 \) and \( n_2 \), and that these are used to form estimates \( N_1 = n_1 - <N_2> \) and \( N_2 = n_2 - <N_2> \) of the true signal returns \( <N_1> \) and \( <N_2> \) where \( <N_2> \) is the sky background count, assumed known. An estimate \( \log_e(\frac{<N_1>}{<N_2>}) \) of the true value of the logarithm difference, \( \log_e(\frac{<N_1>}{<N_2>}) \), is then obtained. The bias, standard deviation, and rejection ratio of the log difference estimates, have been obtained from simulations in which Poisson statistics were assumed for \( n_1 \) and \( n_2 \), and results are exemplified in Figs. 36 and 37. These results were calculated for fixed values of \( \frac{<N_1>}{<N_2>} \) and \( -\log_e(\frac{<N_2>}{<N_1>}) \), where the latter is proportional to the column content of the measured species. In general terms it is seen that the bias can be neglected if \( N_1 > 30 \) when \( <N_2> = 0 \), if \( N_1 > 200 \) for \( \frac{<N_1>}{<N_2>} = 0.1 \), and \( N_1 > 1000 \) for \( \frac{<N_1>}{<N_2>} = 0.01 \). From the single pulse photocounts available from the smaller (10 mW) system of Table 3 and assuming a pulse prf of 6 kHz, we find that sufficient signal will be available for unbiased estimates of column content within measurement periods less than about 2 ms/wavelength at a range of 150 m, 50 ms/wavelength at 3 km, and 1 s/wavelength at 6 km. These times are less than those generally associated with atmospheric fluctuations.

These figures relate to mean returns. With atmospheric fluctuations, the actual returns and the corresponding integration times will vary. However, range-resolved estimates are obtained by forming the difference of neighboring column content estimates and the bias tends to cancel, so it is less of a problem than for column content estimation. Therefore even with the lowest power systems...
Figure 36: Spread in estimates of the log difference $\log_e(\langle N_1 \rangle/\langle N_2 \rangle)$ as function of $\langle N_1 \rangle$ for $\langle N_1 \rangle/\langle N_2 \rangle = 0.1$, and true values $|\log_e(\langle N_2 \rangle/\langle N_1 \rangle)| = 0$ (heavy line), 1.0 (dashed line), and 3.0 (solid line). The spread is characterized by +/- 1 standard deviation, $\sigma$. There is no bias for $\log_e(\langle N_1 \rangle/\langle N_2 \rangle) = 0$ but for the other values the log difference is overestimated at low $\langle N_1 \rangle$. The bias ceases to be apparent and $\sigma$ decreases as $1/\langle N_1 \rangle$ for $\langle N_1 \rangle > 30$.

Figure 37: Bias (solid lines) and reject ratio (dashed lines) for $|\log_e(\langle N_1 \rangle/\langle N_2 \rangle)| = 1.0$ at different background levels: zero background (empty squares), $\langle N_1 \rangle/\langle N_2 \rangle = 0.1$ (+) and 1.0 (X). Most of the bias at low signal levels is due to rejections, but there is some residual (negative) bias when no estimates are rejected.
considered here, the returns should be adequate for following likely atmospheric fluctuations, and the DIAL wavelengths should be switched faster than the fluctuations.

9.3 Simulation results

Figure 38 shows the simulated nighttime performances of both types of direct detection DIAL systems (pulsed and rm-code) at 828 nm assuming the parameters in Table 2 and output powers of ~10 mW (typical of a diode laser with a monolithic amplifier) and ~350 mW (eye-safe limit). The results for daytime conditions are shown in Fig. 39. The simulations show that acceptable performance can theoretically be obtained with a small diode-based DIAL system based on these specifications.

Relative to pulsed detection, the rm-code modulation can be done with less expensive and more readily available cw lasers. However, when compared with the pulsed method, the signal-to-noise ratios (SNRs) from this technique are considerably lower (by a factor of about 16) for the same size laser beams at the eye-safe power limits.

The sources of noise in direct detection lidar measurements are speckle fluctuations, shot noise from the return signal, shot noise from the sky background, and the detector dark current. The standard deviation of the measurement noise, $\sigma$, is given by the square root of the sum of the variances of these terms:

$$\sigma = \sqrt{\text{var}_{\text{signal}} + \text{var}_{\text{speckle}} + \text{var}_{\text{sky}} + \text{var}_{\text{detector}}}$$

where

$$\text{var}_{\text{signal}} = P_s(i) \frac{k\nu}{T},$$

$$\text{var}_{\text{speckle}} = P_s \frac{\rho^2 \tau_{\text{pulse}}}{n_{\text{speck}} \frac{k\nu}{T}},$$

$$\text{var}_{\text{sky}} = P_{\text{sky}} \frac{k\nu}{T},$$

and

$$\text{var}_{\text{detector}} = \frac{\text{NEP}_{\text{det}}^2}{T}.$$ 

Here $T$ is the total measurement time for one range gate for the entire experiment ($T = \frac{r_{\text{step}}}{c} \frac{t_{\text{exp}}}{\text{prf}}$). Because the noise is generated by random processes, the variances are assumed to have a gaussian probability distribution.

At low transmit powers, the noise is primarily due to the sky background, while at higher powers, the shot noise dominates at the lower altitudes. The noise can be reduced by optimizing the receiver acceptance angle. (Fig. 40) Higher SNRs can also be achieved with higher prfs, because the disadvantage of increased background integration is more than compensated by the roughly fourfold
Figure 38. Simulated nighttime performances for direct detection DIAL systems for the parameters in Table 2.

Figure 39. Same as Fig. 36 except for daytime conditions.
Figure 40. Same as Fig. 37, except for various acceptance angles. The minimum acceptance angle to accept all of the light from the maximum range of 6 km is $\theta = 59 \mu$rad.

Figure 41. Simulated heterodyne DIAL for 1-min. integration times and a 6-cm beam radius.
increase in allowed eyesafe average power. A higher prf also results in higher average powers from diode lasers because they are peak power limited.

The simulated performances of some heterodyne DIAL systems are shown in Fig. 41. The beam radius of 6 cm is the minimum radius which permits the beam to be in the near field out to 6 km. For this 6-cm beam radius, the maximum eyesafe mean power is 11 W, however, such a laser would be quite expensive. Increasing the integration time of 1 minute by a factor \( n \) would increase the SNRs by a factor \( \sqrt{n} \).

In a heterodyne system, the noise sources are shot noise and speckle fluctuations:

\[
\sigma = \frac{P}{\sqrt{N}} \left( 1 + \frac{h v}{\tau_{\text{pulse}} P} \right)
\]

where \( P \) is the return power without noise, \( \sigma \) is the standard deviation of the return power, \( h v \) is the energy of one laser photon, and \( \tau_{\text{pulse}} \) is the laser pulse duration. \( N \) is given by \( n_{\text{shot}} \cdot \tau_{\text{range}} / \tau_{\text{pulse}} \) where \( n_{\text{shot}} \) is the number of laser shots, \( \tau_{\text{range}} \) is the time that signal is received from one range. The SNR is near optimum if the return power, \( P \), results from one photon per bandwidth interval, \( B \), and improves by less than a factor of two if \( P \) is further increased. The noise increases substantially if \( P < h v B \) where \( h v \) is the energy of one photon at frequency \( v \). The prf must be adjusted to balance these competing energy terms; for these low power systems, better results are obtained at lower prfs.

9.4 Comparison of rm-code and pulsed direct-detection lidars

Relative to pulsed direct detection, the rm-code modulation can be done with less expensive and more readily available cw lasers. In Fig. 32 it is seen that at 800 nm, a cw beam (used for the rm-code technique) may have 5 times more power than a pulse beam of the same size. However, when compared with the pulsed method, the signal-to-noise ratios (SNRs) from this technique are considerably lower (by a factor of about 16) for the same size laser beams at the eye-safe power limits.

Here we compare the return signals for rm-code and pulsed detection. [Takeuchi et al., 1986] If the transmitted signals have the same average powers for both techniques, then the pulse powers are \( P_{\text{rm}} = \frac{2}{N} P_{\text{pulse}} \) for an \( N \)-bit rm-code which cycles once per pulse of the other technique. For the following discussion, an eye-safety factor, \( f \), is included: \( P_{\text{rm}} = \frac{2 f}{N} P_{\text{pulse}} \). For pulses at 6 kHz and 800 nm, the eye-safety factor is \( f = 5 \). Both techniques measure the same average background signal, but the return signals for the two methods will differ by the factor \( f \).

The signal variances for the two techniques are different. For the direct detection case, the variance in a particular range is a function of the signal only in that range. In the rm-code method, however, the variance in one range depends on the average signal from all ranges as well as the number of code bits, \( N \).
Following Takeuchi et al. [1986] the signal-to-noise ratios (SNRs) for the two methods are:

\[ SNR_{rm}(j) \propto \frac{fG_j}{\sqrt{N} \sqrt{f\bar{G}} + b} \]

and

\[ SNR_{pulse}(j) \propto \frac{G_j}{\sqrt{G_j + b}} \]

Here, \( N \) is the number of bits in the RM code sequence, \( b \) is the background signal, \( G_j \) is the atmospheric return (laser average power * atmospheric response) for the \( j \)th range element, and \( \bar{G} = \frac{1}{N} \sum G_j \) is the average atmospheric return. Because \( SNR_{rm} \) is a function of \( \bar{G} \), it will have a larger range than \( SNR_{pulse} \). When \( b \) is smaller than \( G_j \) and \( \bar{G} \), then \( SNR_{rm} \propto G_j \) while \( SNR_{pulse} \propto \sqrt{G_j} \). The ratio of the SNRs for the two methods is:

\[ \frac{SNR_{rm}(j)}{SNR_{pulse}(j)} = \frac{f \sqrt{G_j/b + 1}}{\sqrt{N} \sqrt{f\bar{G}/b + 1}}. \]

Note that the background signal, \( b \), has been assumed to be constant and accurately measured; if the variance on the measurement of \( b \) is substantial, then the terms "\( b \)" become "\( 2b \)" in these equations.[Takeuchi et al., 1983]

The SNR ratio at a specific range depends both on the return at that range as well as the average return. The SNR ratio is less than 1 in all cases; i.e. the pulsed technique is always better. For the bright sky case, \( \bar{G} < b \), the SNR ratio is nearly linear with respect to atmospheric return:

\[ \frac{SNR_{rm}(j)}{SNR_{pulse}(j)} \approx f \frac{5}{\sqrt{N}} = 0.3, \]

while for nighttime conditions, \( (\bar{G} > b) \), the ratio is:

\[ \frac{SNR_{rm}(j)}{SNR_{pulse}(j)} \approx \frac{\sqrt{f} \sqrt{G_j}}{\sqrt{N} \sqrt{\bar{G}}} = 0.14 \frac{\sqrt{G_j}}{\sqrt{\bar{G}}}. \]

In daylight and with eyesafe conditions at 800 nm (\( f=5 \)), the SNR of the pulsed lidar method is better by a factor of 3 than the rm-code technique. Under dark sky conditions, the pulsed method is even more preferable at all but the lowest altitude ranges.

Several conclusions are reached in this simple evaluation. First, based on eye-safety
considerations, at shorter IR wavelengths, the rm-code technique permits a return signal 5 times larger than that of the pulsed technique. Nevertheless, the SNR of the direct detection always exceeds that of the rm-code technique. Therefore, the rm-code technique is preferable over pulsed direct detection only at wavelengths where the available pulsed lasers have much less power than that from cw lasers and permitted by the eye-safety limits.

9.5 Comparison of heterodyne and direct-detection lidars

The eye-safe power limits permit a higher power signal with a smaller beam at the longer IR wavelengths near 1.5 \( \mu \text{m} \). As can be seen in Figs. 38-41, good measurements can be made with either a direct detection laser system or a heterodyne system. However, because a direct-detection system near 800 nm would have adequate resolution and would be simpler, lower-power and still eye-safe, we recommend that a water vapor DIAL system be designed at these wavelengths.

9.6 Conclusions

The simulations demonstrate that a small, ground-based DIAL system can measure water vapor profiles to 6 km. The best results are certainly obtained if the transmitter can produce power at the eye-safe limit. The SNR can be optimized by adjusting the prf, acceptance angle, and water vapor line strength. For better results, the system could be run with altitude-dependent integration times: short times at low altitudes, and longer times (or bigger range gates) above the boundary layer where there is less signal and less turbulence.

For a given laser, the models will be very useful to optimize the transceiver design. The next step in the design of this water vapor lidar system is to determine the specifications of possible lasers and detectors.

10 System design

10.1 Detectors

The performances of photon-counting detectors versus wavelength is shown in Fig. 42. Besides sensitivity at longer wavelengths, avalanche photodiode (APDs) have a higher quantum efficiency (QE) and are much more compact and rugged than photomultiplier tubes (PMTs). Both systems can have dark count as low as ~20 counts per second. The primary APD available for photon counting is the SPCM-AQ, an actively quenched system which is manufactured by EG&G Canada. Both PMTs and APD have afterpulsing which presents the challenge of extracting the real signal from the detector artifacts.

The disadvantage of the APDs is that the active quenching circuit which removes the strongest afterpulses results in 50 ns of dead time after each received photon. For high light levels, over 1 MHZ or several pulses in a 100-m range, some pulses will be missed and therefore a correction curve is supplied by EG&G based to account for this. To missing pulses in a lidar, a second detector in analog
Figure 42. Performance of photon-counting detectors versus wavelength. (Figure adapted from *Photomultiplier Tubes*, Hamamatsu Catalog No. TPMO0002E01, Hamamatsu Photonics K.K., Japan, 1994.)

Figure 43. Geometric overlap factor for a receiver field-of-view of 200 μrad and laser transmitter divergences of 100, 200, and 400 μrad. (from P. W. Randall, Masters thesis, University of Arizona, 1995.)
mode could be used for the lowest ranges.

Intensified photodiodes (IPDs), which are hybrids of PMTs and APDs are being developed at several companies. IPDs will potentially expand the wavelength range of photon counting and hence direct detection to 1.6 μ. However, at present, IPDs are too early in the development stage to consider for use in the water vapor DIAL system. During the summer of 1996, we may have the opportunity to test a prototype IPD for its noise characteristics.

10.2 Lasers

The design of this DIAL system is constrained by four issues: eye-safety, suitable water vapor lines, detector sensitivity, and potentially available lasers. The laser should be single frequency with a TEM\textsubscript{00} mode, pulse durations between 10 and 200 ns, no prepulse light and a prf of <6 kHz. An all solid-state laser system has the advantages in terms of simplicity, size, cost, and stability. A broadly tunable laser permits access to the best water vapor lines and also can be switched to different lines as the atmospheric conditions change. All of the laser systems considered are somewhat experimental and require development.

A diode-based laser system is preferred because diode lasers are small, simple, rugged, very energy-efficient (~40% QE), and inexpensive to produce in large quantities. Advanced fabrication techniques have resulted in tunable, single-frequency diodes for applications requiring high spectral quality as well as high power laser diode arrays for efficient laser pumping. For example, a Nd:YAG laser pumped with a laser diode array would have a QE of ~5% while one which is pumped with a broadband flashlamp would have a QE of ~1%. The disadvantages of diode lasers are large beam divergence and astigmatism which can be mostly corrected, peak power limitations, and occasional catastrophic failures. Extremely short excited-state lifetimes of the semiconductor materials result in fast response times as well as the peak power limit.

For the purposes of compactness and simplicity, an all laser-diode system is appealing. At this time, however, the pulsed power from single-frequency diode lasers is unacceptably low for the DIAL system. However, diode lasers can serve as a single-frequency source either for amplification or for injection seeding. Microchip lasers are also single-frequency, but have tuning ranges of <30 GHz and exist at only a few wavelengths which are not suitable for this water vapor lidar.\cite{Zayhowski and Dill} Extended cavity diode lasers are tunable, stable, and single frequency. Distributed feedback (DFB) lasers and distributed Bragg reflector (DBR) lasers are smaller, less vibration sensitive and less expensive than extended cavity lasers, but are only available at select wavelengths. Four diode-based laser candidates are considered for a pulsed incoherent DIAL system: (1) an 828-nm pulsed diode laser with flared amplifier, (2) a 940-nm cw DBR diode laser with pulsed fiber amplifier, (3) a tunable diode-pumped Cr:LiSAF or related laser, and (4) a tunable Ti:sapphire laser pumped with a diode-pumped Nd:YAG laser.

10.2.1 All-semiconductor laser

A simple, rugged system would be a cw diode seed laser fiber-coupled to a flared diode amplifier. The disadvantage of semiconductor materials is that they are peak-power limited because they have
very short excited-state lifetimes. For example, a commercially available single-frequency diode laser, provides <2 W of peak power near 828 nm. After amplification with a kHz-rate pulsed diode laser flared amplifier, [Krainak et al.] the average power is ~10 mW; this is 10 times less power than permitted by eye-safe limits for a 10-cm radius beam.

Increased transmitted power can be obtained by splitting the diode laser beam among several parallel flared amplifiers. The amplified beams could then be directly transmitted or first recombined in pairs with polarizing beam splitters. Experimental flared amplifiers, fabricated by SDL, Inc., have been used to produce >5W of average power and >10W of peak optical power. [Mehuys, Goldberg, and Welch] These may be soon be available commercially for use in a DIAL system.

10.2.2 Diode laser with fiber amplifier

An option for the lidar transmitter with potentially higher output power is a tunable 920-nm DBR laser with a Nd-doped fiber amplifier. There are many design challenges in both the DBR and the fiber development. However, with recent fiber technology this system might produce output powers near the eye-safe limit.

Researchers at the University of Illinois have been working on a design for a ~940-nm transmitter using a DBR laser and a fiber amplifier which could be used for a lidar for high altitude water vapor studies. [Papen et al.] For the ground-based lidar we would want a similar system, but at weaker water vapor absorption line, probably around 920 nm. This transmitter would have a similar design, but the materials will be modified in both the DBR and the fiber.

The Coleman group has fabricated DBR lasers at 1.06 μm with good tuning repeatability, between 10 and 50 mW of output power, ~40 dB of side-mode suppression and extraordinarily narrow linewidths as low as 40 kHz (as measured by self-heterodyning). There are still some unresolved problems with the 1-μm lasers, most notably regarding chirp. At present, the group is trying to controllably fabricate DBR lasers at shorter wavelengths.

It is likely that the problem of chirp may not be correctable and that the DBR lasers will require external modulation. Whether this should be done with a nanofabricated device on the same wafer as the laser or with an acousto-optic modulator remains to be investigated. If chirp while tuning cannot be removed, then either some pulses will need to be neglected, or two separate lasers should be used. Also, there will be the option of fabricating the two lasers together with a waveguide.

The proposed amplifier for a diode laser / fiber amplifier system would be made of a Nd-doped fiber which has a tuning range with a FWHM of ~30 to 40 nm. The fibers can be pumped with laser diodes at 808 nm, one of the most widely available diode wavelengths. The doping of the fiber with Nd, Al, and P, affects the lifetime, gain and wavelength of the amplifier. For a three level system such as these Nd lines, it is crucial to select the correct length of fiber relative to the doping concentration. The system should be saturated throughout the length of the fiber. If the fiber is too short then pump light is wasted, and if it is too long, then laser light will be reabsorbed.
With single-mode fibers, the amplifiers may be 5% efficient. New technology using double-clad fibers and V-grooves to admit the pump light permit efficiencies greater than 50%.[Ripin and Goldberg] Although these high power amplifiers would be more expensive, they would result in output powers near the eye-safe limit.

10.2.3 Other solid state laser options

Two other options for tunable, kHz-rate all solid state lasers are Cr:LiSAF and Ti:sapphire, both of which could be injection-seeded near 830 nm and operated near the eye-safe power limit. The Cr:LiSAF laser would be diode-pumped at 670 nm while the Ti:sapphire laser would be pumped by a diode-pumped Nd:YAG laser.

A mJ-pulse, diode-pumped Cr:LiSAF laser would be challenging to build due to the short excited-state lifetime and thermal problems of the crystal. [e.g., Zenzie, Finch, and Moulton] Because the excited state lifetime of Cr:LiSAF is 67 \mu s, the optimum pulse repetition rate is about 15 kHz. So to extract the maximum energy from the system, it would be run at the highest prf feasible for the DIAL system, about 10 kHz. Possible solutions to the thermal problems include careful design with a loose focus inside the crystal or the use of a related Cr-doped crystal with a slightly different gain spectrum and better thermal loading than Cr:LiSAF.

A kHz-rate Ti:sapphire laser might be slightly easier to develop than a Cr:LiSAF laser because the material has been better characterized and produced with higher quality.[e.g., Hamilton] Researchers at Sandia National Lab have injection seeded a Ti:sapphire laser with extended cavity laser diodes at a prf of 30 Hz for use in an airborne (non-eyesafe) water vapor DIAL system. [Henson et al.] The high cost of a diode-pumped Nd:YAG laser required to pump the crystal is prohibitive for this project, however.

A passively Q-switched Er-doped laser and fiber amplifier near 1.5 \mu could also be used for a lidar. This system would require the added complexity of heterodyning, though, because there are no photon-counting detectors at wavelengths beyond 1 \mu.

10.3 Other system components

An essential element in this system is the narrowband receiver filter which limits the daytime sky background. This can be done by passing the return beam through an etalon and an interference filter. As an example, pressure-controlled etalon manufactured by Melles Griot has a free spectral range of 300 GHz and a linewidth of \sim 30 GHz. It should be noted that although the etalon is rated for 80% transmission for a laser beam, the transmission will be only about 40% in the lidar system because the return beam has increased divergence. The etalon must followed by a temperature controlled interference filter to remove all but one of the lines passed by the etalon. An example interference filter is made by Andover and has a bandpass of 3 \AA and 35% transmission. The filter should follow the etalon to minimize the loss in beam quality prior to the etalon.

Several design options for a generic lidar transmitter have been examined in detail by Paul Randall at the University of Arizona (contractor) and presented in his Masters thesis. [Randall] In particular,
two designs are considered for minimizing the laser diode beam divergence and for expanding the beam to ensure eye-safety. The first design utilizes a single large aspherical lens, while the second design requires a beam collimator, a beam circularizer, and a beam expander.

The design options were evaluated using the optical ray tracing program, ZEMAX. [ZEMAX] The diode's asymmetrical divergence and astigmatism were modeled using a point source and a paraxial XY surface. The design with the single aspherical lens is theoretical possible, but would be extremely expensive in practice. For the multi-component design, a glass asphere is found to be the best collimator. It has less power loss and a somewhat lower cost than a compound spherical lens, and half the divergence of a GRIN lens. It is found that there is little tolerance on the spacing of the optics in any of the designs. (Table 4) Asymmetry correction is performed with an anamorphic prism pair. Beam expansion can be done with a lens or various mirror pairs including either a spherical mirror or an off-axis parabola mirror. If the mirrors are aluminum and coated with Si₃O₂ to be highly abrasion resistant, then there will be about a 20% loss in transmission. Other coatings will have substantially lower losses. The off-axis option avoids spherical aberrations, but is much more expensive.

The above simulations were performed for a coaxial system. For daytime operation of a non-coaxial system, these studies find that the transmitted beam should be as collimated as possible, and the receiver field-of-view should no more than 200 μrad. When just the geometric overlap is considered for a receiver field-of-view of 200 μrad, the optimum laser transmitter divergence is about 100 μrad. (Fig. 43) These simulations demonstrate that a small divergence transmitter beam from a diode laser system is feasible with carefully selected optics.

<table>
<thead>
<tr>
<th>transmitter type</th>
<th>separation</th>
<th>nominal separation (mm)</th>
<th>half-angle divergence goal (μrad)</th>
<th>±tolerance (μm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lens expander</td>
<td>laser diode to asphere</td>
<td>4.8</td>
<td>50</td>
<td>17</td>
</tr>
<tr>
<td>system</td>
<td>lens separation</td>
<td>269.4</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>mirror expander</td>
<td>laser diode to asphere</td>
<td>4.8</td>
<td>50</td>
<td>47</td>
</tr>
<tr>
<td>system</td>
<td>mirror separation</td>
<td>240.0</td>
<td>100</td>
<td>90</td>
</tr>
<tr>
<td>single lens</td>
<td>laser diode to asphere</td>
<td>250.0</td>
<td>50</td>
<td>9</td>
</tr>
<tr>
<td>system</td>
<td>to asphere</td>
<td></td>
<td>100</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 4. Lidar transmitter tolerance data summary. (From P. W. Randall, Masters thesis, University of Arizona, 1995.)

11 Laboratory studies

A variety of laboratory tests were conducted by Laurie Kennard at the University of Arizona (contractor) to examine the feasibility of using laser diodes in a water vapor DIAL system. [Kennard] The spectral shifts, mode hops, and output power of near-infrared laser diodes were characterized as functions of temperature and current. Other work involves a comparison of techniques for measuring the central wavelength and spectral linewidth of a pulse modulated laser diode.
Experiments were performed on an SDL-5311 laser diode, manufactured by SDL, to spectral shifts, mode hops, and output power of near-infrared laser diodes were characterized as functions of temperature and current. Mode hops occur when either the temperature or the current are varied. The mode hops due to temperature changes are not monotonic, and involved shifts of more than 1 nm for a temperature change of 1 C. The linewidth is constant for the different wavelengths.

Three spectral instruments were studied: Fabry-Perot interferometer, a grating monochromator, and a Michelson interferometer. A comparison was done of the capability of these instruments to measure the central wavelength and spectral linewidth of a pulse modulated laser diode. The complexity of design and the ease of operation were also considered. A Fabry-Perot interferometer can measure both the central wavelength and the spectral linewidth with great accuracy. A grating monochromator can measure the central wavelength with relative simplicity and moderate accuracy. A Michelson interferometer can measure an accurate spectral linewidth after time-averaging calculations.

The Fabry-Perot interferometer considered is the NASA/LASE wavemeter which has the stringent requirement that the wavelength to be predetermined to within 170 picometers (pm).[Goad et al.] As a three stage system, this allows for high accuracy determination of the central wavelength (±0.1 pm) and a bandwidth measurement of about 20 GHz.

12 Conclusion

In conclusion, the simulations indicate that a small DIAL system is feasible for water vapor measurements in the lower troposphere. A pulsed lidar gives better performance than a comparable system which uses pseudo-random code modulation. Both direct detection and heterodyning would give adequate results for a ground-based DIAL system, but direct detection is preferable for this ground-based system due to its lower cost and simplicity. Our analysis shows that there will not be problems with bias and atmospheric variability if the DIAL wavelengths are switched rapidly enough.

Essential elements needed to obtain acceptable DIAL performance are a frequency-stable source operating at a high prf and a narrowband receiver to minimize background light. A diode-based system is desired because it is compact, rugged, energy-efficient and low-cost. Two options for the lidar source are to use a DBR laser amplified by either a diode flared amplifier or a rare-earth doped fiber. Other system components have been examined including filters, wavemeters and collimating optics, and laboratory experiments were performed to examine the stability, tunability, and spectral characteristics of diode lasers. Because diode laser technology is advancing rapidly, there is considerable promise that a compact, eye-safe water vapor DIAL system can now be built with good measurement capabilities.
References for Part II


