An Automated Method for the Detection and Extraction of H I Self-Absorption in High-Resolution 21cm Line Surveys

Steven J. Gibson[1], A. Russell Taylor[1], Lloyd A. Higgs[2], Christopher M. Brunt[3], and Peter E. Dewdney[2]

ABSTRACT

We describe algorithms that detect 21cm line H I self-absorption (HISA) in large data sets and extract it for analysis. Our search method identifies HISA as spatially and spectrally confined dark H I features that appear as negative residuals after removing larger-scale emission components with a modified CLEAN algorithm. Adjacent HISA volume-pixels (voxels) are grouped into features in (ℓ, b, v) space, and the H I brightness of voxels outside the 3-D feature boundaries is smoothly interpolated to estimate the absorption amplitude and the unabsorbed H I emission brightness. The reliability and completeness of our HISA detection scheme have been tested extensively with model data. We detect most features over a wide range of sizes, linewidths, amplitudes, and background levels, with poor detection only where the absorption brightness temperature amplitude is weak, the absorption scale approaches that of the correlated noise, or the background level is too faint for HISA to be distinguished reliably from emission gaps. False detection rates are very low in all parts of the parameter space except at sizes and amplitudes approaching those of noise fluctuations. Absorption measurement biases introduced by the method are generally small and appear to arise from cases of incomplete HISA detection. This paper is the third in a series examining HISA at high angular resolution. A companion paper (Paper II) uses our HISA search and extraction method to investigate the cold atomic gas distribution in the Canadian Galactic Plane Survey.

Subject headings: radiative transfer — methods: analytical — techniques: image processing — surveys — ISM: clouds — ISM: structure

---

[1]Dept. of Physics & Astronomy, University of Calgary, 2500 University Drive N.W., Calgary, Alberta T2N 1N4, Canada; gibson@ras.ucalgary.ca; russ@ras.ucalgary.ca

[2]Dominion Radio Astrophysical Observatory, Box 248, Penticton, British Columbia V2A 6K3, Canada; Lloyd.Higgs@nrc.ca; Peter.Dewdney@nrc.ca

[3]Department of Astronomy, LGRT 632, University of Massachusetts, 710 North Pleasant Street, Amherst, MA 01003; brunt@roobarb.astro.umass.edu

1. Introduction

The 21cm line of neutral atomic hydrogen (H I) is a key probe of the Galactic interstellar medium. Although the cold ($T \lesssim 100$ K) gas distribution is difficult to map in H I emission, H I self-absorption (HISA) allows cold foreground gas to be distinguished from warmer background gas at the same radial velocity (Gibson 2002). Until recently, HISA has been studied in limited low-resolution maps (e.g., Baker & Burton 1979; Bania & Lockman 1984), or in a few isolated objects at higher resolution (e.g., van der Werf, Goss, & Vanden Bout 1988; Feldt 1993), but no detailed, systematic surveys have been made. High resolution allows feature structure to be studied and unabsorbed background brightness to be estimated accurately. Coverage of a wide area enables an unbiased look at the HISA population, e.g., without the a priori expectation that HISA is found only in molecular clouds.

High-resolution, wide-area HISA surveys have now become possible with the advent of several major H I synthesis surveys: the Canadian Galactic Plane Survey (CGPS; Taylor et al. 2003), the Southern Galactic Plane Survey (SGPS; McClure-Griffiths et al. 2001), and the VLA Galactic Plane Survey (VGPS; Taylor et al. 2002). Past HISA studies have identified absorption features by eye, but this approach is no longer adequate. The very richness of the synthesis survey data sets requires that they be analyzed in a rigorous, repeatable manner. We have therefore designed automated algorithms to identify and extract HISA features from H I longitude-latitude-velocity (ℓ, b, v) data cubes.

In this paper, we describe our HISA search and extraction algorithms. We also explain how we have tested our software with model data to determine its reliability under a range of different conditions. Large surveys are playing an increasingly significant role in modern astrophysics, and it is essential that their underlying methods are understood so their results can be interpreted properly. Following criteria established in Gibson et al. (2000; hereafter Paper I), our HISA search software seeks finely-structured dark features against bright backgrounds that cannot be confused with simple gaps in H I emission. Although its parameters are optimized to identify HISA in the CGPS, the software is easily adapted to work with other surveys (e.g., the VGPS: Gibson et al. 2004).

The CGPS uses a hexagonal grid of full-synthesis fields with single-dish observations to enable the detection of all scales of H I structure down to the synthesized beam. The CGPS H I data have a $58'' \times 58''\mathrm{cosec}(\delta)$ beam, 0.824 $\mathrm{km\,s^{-1}}$ velocity sampling, and a field-center noise of $T_{rms} \sim 3$ K in empty channels; T_{rms} doubles when the $107'$ primary beam is filled with 100 K emission, and it can be up to 60% greater between field centers. The initial phase of the survey mapped a $73^\circ \times 9^\circ$ region along the Galactic plane with longitudes $74.2^\circ \leq \ell \leq 147.3^\circ$ and latitudes $-3.6^\circ \leq b \leq +5.6^\circ$ ($+33.9^\circ \leq \delta \leq +68.4^\circ$), and extensions in both ℓ and b have followed.

Below, we describe our method of HISA identification and extraction at some length (§2) and evaluate the method's performance with models (§3). A companion paper presents the results of our HISA search of the $73^\circ \times 9^\circ$ Phase I CGPS (Gibson et al. 2005; hereafter Paper II). Subsequent papers in this series will apply the HISA search to other data sets.

2. Feature Extraction

2.1. Identification Strategy

2.1.1. Criteria

Many HISA features are apparent to the eye (e.g., see Figs. 4 - 8), but a complete visual search is unlikely to be uniform, repeatable, or thorough, and it is also impractical given the sheer volume and complexity of the CGPS data. Thus, an automated search is needed. The search algorithm should find features meeting simple criteria that can be confirmed by eye, but it should also be tested with model data to evaluate its performance quantitatively (§3).

The nature and appearance of HISA dictate how it can be identified. First, while the cold H I from which it arises can have any extent, no HISA feature can exceed the (ℓ, b, v) boundaries of its bright background H I emission, or it ceases to be absorption. Second, HISA must have different (ℓ, b, v) structure than the background H I for it to be distinguishable from background fluctuations. We choose to search for HISA that is more finely-structured than the background H I, since this is consistent with the first constraint, most CGPS HISA that can be visually identified is of this nature, and the exceptions (e.g., Knee & Brunt 2001; Kerton 2005) are difficult to identify algorithmically.

We seek H I features that can only be explained as HISA. We prefer this conservative approach over the alternative of including significant false detections in our survey sample. As given in Paper I, our conditions for distinguishing HISA from simple gaps in H I emission are: (1) narrower line widths than most observed emission features; (2) steeper line wings; (3) more small-scale angular structure; and (4) a minimum H I emission background level. The first two conditions are related for Gaussian line profiles, since these have line wing slopes proportional to amplitude over width, but real HISA need not be Gaussian. The last condition excludes the finely-structured H I emission gaps that are common at interarm velocities in the outer Galaxy, where smooth, bright H I backgrounds are often absent. These four criteria exclude HISA on larger angular and velocity scales or against weaker backgrounds, but they are adequate for capturing most visible features. We do not require the extra condition of molecular line emission to confirm HISA features (e.g., Knapp 1974),

since many HISA features are visible without ^{12}CO emission in the CGPS (Paper I; Gibson 2002) and in other surveys (e.g., Peters & Bash 1987).

2.1.2. Algorithms

We tried and rejected many different methods before selecting the algorithm described in this paper. Discarded techniques include various derivative measures to detect sharp edges, spatial and spectral curvature tests to look for dips, Fourier and wavelet filtering methods, and unsharp masking. Most of these were successful in locating the strongest features, but few were robust against noise, and many also produced large numbers of artifacts and false detections. The latter were especially frequent in methods that used only spectral or spatial searches rather than both combined.

Our chosen method is based on a variant of the CLEAN algorithm (Högbom 1974) developed by Steer, Dewdney, & Ito (1984) (hereafter the SDI CLEAN). We remove large-scale spectral and spatial emission structures from the H I data iteratively and flag the small-scale negative residuals as self-absorption features. For computational efficiency, these operations are carried out separately on the spectrum at each spatial position (ℓ, b) and on the channel map at each radial velocity (v) in the data cube, and the results are combined afterward. The identified HISA is then subtracted from the H I cube, and the whole process is repeated until significant HISA can no longer be found; such iteration allows features larger than the chosen CLEAN scales to be mapped.

The two spectral and spatial search algorithms are described below. Each has been tuned to find as much visually identifiable HISA as possible while minimizing false detections. The latter are further reduced by subsequently requiring the HISA at any 3-D position (ℓ, b, v) to be detected by both searches (§2.4). The two algorithms were tuned by visually comparing the search output against the observed H I for many different parameter value combinations, using a range of different HISA features with different H I emission backgrounds in the CGPS data. The parameters that yielded the most complete HISA detections with the fewest false detections were used in the model-based search performance evaluations (§3) and in the CGPS HISA survey (Paper II).

In the following discussions, the Galactic coordinate variables (ℓ, b, v) are replaced by their pixel coordinate analogs (i, j, k). Adopted values for the search parameters are given in square brackets [ ]. These give the best performance for a HISA search of CGPS data, but they may not be universal. In particular, the best filter scales and minimum background level may differ for HISA searches elsewhere in the Galaxy.

2.2. Spectral Search

At each spatial position (i, j), the spectral search algorithm builds an approximation of the "unabsorbed spectrum" $U(k)$ that would be observed if no HISA were present. The algorithm assumes that $U(k)$ can be constructed from Gaussian functions of a characteristic width W_{char} that is narrower than the dominant emission features but broader than the width of any expected HISA feature. Any channels in which the observed spectrum $O(k)$ deviates significantly negatively from $U(k)$ are flagged as possible HISA.

The iterative procedure used to derive $U(k)$ is a modification of the SDI CLEAN. $U(k)$ is initially set to zero, and the "residual spectrum" $R(k)$ is set equal to $S(k)$, a smoothed version of $O(k)$. Smoothed data are used to improve the signal-to-noise for the CLEANing process. $S(k)$ is a *spatial* average of $N \times N$ pixels centered at (i, j), i.e., the average of $O(i', j', k)$ intensities where $|i' - i| \leq (N-1)/2$ and $|j' - j| \leq (N-1)/2$ [and $N = 7$ pixels $= 2.1'$]. Independent of this spatial averaging, the *spectral rms* noise σ_{obs} in $O(k)$ is computed as the lowest of three *rms* noise measures over equal thirds of $O(k)$. In the CLEAN loop, the following steps are performed:

1. If R_{max}, the peak value of $R(k)$, is less than a preset fraction [3%] of the peak value of $S(k)$, the iteration ceases.

2. For any channel k where $R(k)$ exceeds a given clip level [0.8] $\times R_{max}$, a "correction spectrum" $C(k)$ is set to a preset gain [0.25] $\times R(k)$; elsewhere, $C(k)$ is set to zero.

3. $C(k)$ is convolved with a Gaussian whose full width at half maximum (FWHM) is W_{char} [8 $\mathrm{km\,s^{-1}}$], and the resulting spectrum is added to $U(k)$.

4. The new residual spectrum is set to $R(k) = S(k) - U(k)$, and σ_{pos}, the *rms* of all *positive* values of $R(k)$ is computed over all channels. As $U(k)$ approaches $S(k)$, σ_{pos} decreases; if $\sigma_{pos} < \sigma_{obs}$, the iteration is terminated.

5. If the iteration has not terminated due to one of the above convergence criteria, and a maximum number of loops [1000] has not been reached, steps 1-5 are repeated.

After the CLEAN loop is completed, adjacent channels where $R(k) < \sigma_{pos} \times$ a factor F [-2.0] are grouped into "segments" of suspected HISA. In each segment, $S(k)$ and $U(k)$ are evaluated at the channel k_{min} where $R(k)$ has a local minimum. If $S(k_{min}) < 2F\sigma_{pos}$, the segment is rejected as a likely noise feature. If $U(k_{min}) < T_{crit}$, where T_{crit} is a preset brightness level [30 K], the segment is rejected as having an insufficiently bright background

to identify HISA clearly. A second, more conservative $T_{crit}{}'$ [70 K] is applied later when the spatial and spectral search results are combined (§2.4).

A Gaussian profile is fitted to the absorption magnitude spectrum $R(k)$ of each remaining channel segment. Real HISA line profiles may not be Gaussian, but this shape is assumed for simplicity. For computational speed, the central channel of the Gaussian is fixed at k_{min}, and the FWHM is fixed at one of two values, W_{narrow} [2.0 $\mathrm{km\,s^{-1}}$] or W_{broad} [4.0 $\mathrm{km\,s^{-1}}$], to capture HISA with a range of widths. Model tests (§3) show that many features outside this range are also detected. In each fit, a sloping linear base level is derived along with the Gaussian amplitude A and the standard deviation σ_{fit} of the fit. A fit is rejected as statistically unreliable if $A/\sigma_{fit} < D$ [2.0] or $A/\sigma_{pos} < D$. It can also be rejected if $O(k)$ lacks a morphological "dip" at k_{min}. This is determined with a filter function that returns a value of 1.0 for a dip between two equal peaks, 0.5 for a "dip" that drops only to the level of the adjacent spectral data on one side of k_{min}, and 0.0 on a linearly rising or falling spectrum. Fits that return a value below the chosen threshhold [0.6] are rejected. This filter inhibits the detection of HISA on the edges of emission features that are steeper than W_{char} would allow; otherwise, significant false HISA detections result. For accepted fits, the channels in a *narrow* or *broad* HISA spectrum (both initially zero) are set equal to the fitted Gaussian if the amplitude exceeds a given fraction [5%] of A. Then, a "detected" HISA spectrum is created that consists of the maximum value in each channel from the *narrow* and *broad* Gaussian fits to the HISA line profile, to ensure full detection of the feature. In the case where the *broad* line wings do not correspond to real HISA in a narrow feature, these will not be detected by the spatial search and will be removed at a later stage of analysis.

Finally, when the HISA amplitudes have all been computed, these are spatially smoothed [with a $1.5'$ beam] to join together groups of flagged "flecks" into more coherent features, and those that are sufficiently weak and isolated are culled if their amplitude falls below a specified threshhold [2 K]. An additional cosmetic improvement is made by excluding strong HICA from the set of spectrally-identified HISA. This is done by dropping any sight lines from the search that contain channels whose continuum-subtracted line brightness is significantly negative, i.e., if $O(k) < -6\,\sigma_{obs}$. Weaker HICA will survive this filter to contaminate the set of detected HISA. Such contamination is difficult to remove in a way that leaves the "pure" HISA in the same sight lines intact.

To illustrate the algorithm, we plot the H I spectrum of Paper I's Perseus HISA Globule ($\ell = 139.635°$, $b = 1.185°$) in Figure 1 and several stages of the algorithm's analysis in Figure 2. These H I data supersede those used in Paper I, which contained a flaw that had no serious impact on the results. Appendix A gives further details.

In Figure 2, the estimation of $U(k)$ converged when the peak residual became less than

3% of the initial spectral peak, giving $\sigma_{pos} = 3.6$ K (the dashed line indicates a negative deviation of twice this value). Suspected HISA was identified in eight channel segments. One of these (4) had no counterpart dip in $O(k)$. Six others were rejected because $U(k) <$ 70 K. For simplicity, we used $T_{crit} = 70$ K here to show what would survive the ultimate $T_{crit}{}' = 70$ K filter. In the remaining segment (5), channels in which the "merged" HISA spectrum (maximum of the narrow and broad HISA spectra) is non-zero were then flagged as having "detected" HISA. Note that, because of the initial smoothing, the detected HISA has a smaller amplitude than in Figure 1. The full HISA amplitude is recovered when the spectral and spatial search results are merged (§2.4).

2.3. Spatial Search

The spatial search algorithm is similar in principle to the spectral search, although it does not attempt to fit absorption features with Gaussian shapes, nor does it require that they satisfy a morphological "dip" filter. It begins by estimating the unabsorbed brightness distribution $U(i, j)$ in a given spectral channel k. The algorithm assumes that $U(i, j)$ can be constructed from two-dimensional circular Gaussian components of a characteristic width G_{char} that is narrow enough to represent most H I emission structure but broader than any expected HISA features. Clearly the choice of G_{char} limits the angular size of HISA features that will be detected, although this can be alleviated with repeated searches.

The iterative procedure used to derive $U(i, j)$ is again a modification of the SDI CLEAN. $U(i, j)$ is initially set to zero, and the residual map $R(i, j)$ is set equal to $S(i, j)$, a spatially smoothed copy of the observed channel map $O(i, j)$. Use of $S(i, j)$, computed as an $N \times N$ pixel average of $O(i, j)$ [with $N = 15$ pixels $= 4.5'$], improves the CLEAN convergence. On a larger angular scale [$20'$], an estimate of the typical *rms* noise σ_{obs} in $O(i, j)$ and its gross variation across the channel map are derived in a manner similar to the spectral *rms* noise in §2.2. From this, the average *rms* noise in $S(i, j)$, termed σ_{sm}, and its variation over the map are deduced. In the CLEAN loop, the following steps are performed:

1. R_{max}, the Mth [10th] highest value of $R(i, j)$, is found. This is chosen rather than the peak value so that the iteration process is not dominated by one noisy pixel. If R_{max} is less than a preset fraction [3%] of the peak value of $S(i, j)$, the iteration ceases.

2. For any pixel (i, j) where $R(i, j)$ exceeds a given clip level [0.5] $\times R_{max}$, a correction map $C(i, j)$ is set to a preset gain [0.25] $\times R(i, j)$; elsewhere, $C(i, j)$ is set to zero.

3. $C(i, j)$ is convolved with a 2-D Gaussian whose FWHM is G_{char} [$20'$], and the resulting image is added to $U(i, j)$.

4. The new residual map is set to $R(i,j) = S(i,j) - U(i,j)$, and the *rms* value σ_{pos} of the *positive* values of $R(i,j)$ is computed. In deriving σ_{pos}, allowance is made for the fact that the noise may vary across the image by applying suitable weights to the values of $R(i,j)$. If $\sigma_{pos} < \sigma_{sm}$, the iteration is terminated.

5. If the iteration has not terminated due to one of the above convergence criteria, and a preset maximum number of loops [1000] has not been reached, steps 1-5 are repeated.

After the CLEAN loop is completed, all pixels where $R(i,j) < \sigma_{pos}(i,j) \times$ a factor F [-2.0] are noted, as are those where $O(i,j) - U(i,j) < F\sigma_{obs}(i,j)$. A map of "suspected HISA" is set equal to $U(i,j) - O(i,j)$ for all pixels (i,j) where either condition is met and zero elsewhere. This map is then filtered to remove pixels with amplitudes less than a specified cutoff [4 K], as well as those for which $U(i,j) < T_{crit}$ [30 K]. Lastly, as in the spectral algorithm, the suspected HISA map is smoothed [with a $1.5'$ beam] to improve feature coherence, and a final cull is made of smoothed amplitudes below a lower threshhold [2 K]; surviving features are deemed "detected HISA".

The spatial search is illustrated in Figure 3, with longitude profiles taken through the Perseus HISA Globule position ($b = 1.185°$, $v_{LSR} = -41.04$ km s^{-1}) from one channel map at different stages of processing. The determination of $U(i,j)$ took 114 iterations, ending when σ_{pos} became less than $\sigma_{sm} = 2.16$ K.

2.4. HISA Amplitude Estimation

2.4.1. General Approach

The physical properties of the absorbing gas cannot be understood without knowing the HISA brightness temperature amplitude $\Delta T \equiv T_{ON} - T_U$, where $T_{ON}(\ell,b,v) \equiv O(i,j,k)$ is the observed brightness on the HISA feature, and $T_U(\ell,b,v) \equiv U(i,j,k)$ is the unabsorbed emission that would be measured if no HISA were present. Since only T_{ON} is directly observed at the HISA position, T_U must be estimated from T_{OFF}, the H I brightness off the HISA feature in space and/or velocity. For clarity, we note that T_U represents *all* of the unabsorbed emission along the line of sight at radial velocity v. The emission from behind the HISA feature that is subject to absorption is $p \cdot T_U$, where $0 < p \leq 1$; the exact value of p depends upon the sight-line geometry (see Paper I).

Several means of estimating T_U have been used in past studies. In the spectral domain, the velocity edges of a HISA feature can be fitted with straight lines (Hasegawa et al. 1983; Montgomery et al. 1995) or more complex functions (Knapp 1974; McCutcheon et al. 1978; Li

& Goldsmith 2003; Kavars et al. 2003) to estimate T_U at intervening velocities. In the spatial domain, the H I brightness at positions adjacent to the HISA feature can be used directly as T_U (Paper I; Minter et al. 2001) or as anchor points for spatial fits across the feature (Feldt 1993; Kavars et al. 2003). A variant on this approach assumes HISA is sufficiently diluted in the broad beam of a single dish telescope to use the single dish spectrum at the feature position as T_U (van der Werf et al. 1988).

Our approach is more general. We group the HISA volume-pixels (voxels) identified in §§2.2-2.3 into contiguous 3-D features in the spectral line cube. For each feature, we estimate T_U by interpolating the T_{OFF} values of the non-HISA voxels that border the feature in (ℓ, b, v) space. The interpolation uses a 3-D Gaussian weighting function to ensure smoothness on the scale of the feature. Specifically, at each position (ℓ, b, v) within the HISA feature,

$$T_U(\ell, b, v) \;=\; \frac{\sum_{n=1}^{N_{OFF}} w_n \cdot T_{OFF}(\ell'_n, b'_n, v'_n)}{\sum_{n=1}^{N_{OFF}} w_n} \quad , \tag{1}$$

where n indexes the list of N_{OFF} off-HISA voxels with coordinates (ℓ'_n, b'_n, v'_n), and the weight w_n is given by

$$w_n \;=\; \exp\left[-\frac{1}{2}\left(\frac{\ell'_n - \ell}{\sigma_\ell}\right)^2 - \frac{1}{2}\left(\frac{b'_n - b}{\sigma_b}\right)^2 - \frac{1}{2}\left(\frac{v'_n - v}{\sigma_v}\right)^2\right] \quad . \tag{2}$$

The Gaussian dispersions $(\sigma_\ell, \sigma_b, \sigma_v)$ are set so that each FWHM $(= \sigma \cdot \sqrt{8 \ln 2})$ is half the maximum length of any contiguous row of HISA voxels in that dimension, with FWHM lower limits of $1.2'$ and $3.3\,\mathrm{km\,s^{-1}}$ and upper limits of $20'$ and $8\ \mathrm{km\,s^{-1}}$, the HISA search CLEAN scales. The result is somewhat similar to that of 1-D spectral fitting methods, but its structure is constrained by all three dimensions of the H I data. This method yields T_U and ΔT estimates superior to those of our separate spectral and spatial searches.

2.4.2. Filtering

The T_U estimation algorithm considers two confidence levels of HISA. First, a *union* filter requiring a HISA identification from either the spectral or spatial search is applied. This filter includes nearly every HISA feature that the eye can detect, as well as many non-

HISA features that are discarded later. All accepted voxels are grouped into tentative HISA features and interpolated over to obtain $T_{_U}$, which is subtracted from the unsmoothed H I data to get ΔT with the full CGPS angular resolution. Then, an *intersection* filter requiring HISA identification in *both* spectral and spatial searches is applied. Voxels not satisfying this filter are unflagged as HISA, and their $T_{_U}$ is reset to the observed H I brightness. Computing $T_{_U}$ and ΔT for a union voxel set and applying an intersection filter afterward ensures that (1) only the most likely HISA features survive, and (2) any "penumbral" contamination from undetected HISA in their $T_{_{OFF}}$ voxels is minimized; otherwise, $T_{_U}$ and $|\Delta T|$ could be significantly underestimated. An alternative is to interpolate only over HISA satisfying the intersection filter with all union-filter voxels dropped from the $T_{_{OFF}}$ ensemble, but this frequently leaves too few edge voxels for a robust $T_{_U}$ estimate.

Three additional filters are applied with the intersection filter. Voxels with $\Delta T \geq 0$ are discarded, as are those in the noisy peripheries of the survey and those for which $T_{_U} < T_{crit}{}'$ [70 K], a stricter value than the previous T_{crit} [30 K] of §§2.2-2.3. The peripheral culling rejects HISA voxels with CGPS field mosaic weights $w_m < 0.382$, the lowest weight that occurs between synthesis field centers. Since $w_m \propto \sigma_{noise}^{-2}$, this allows a maximum noise of 1.618 times the field center value (see Taylor et al. 2003), which is typically $5 - 7$ K for the $T_{_U} \sim 70 - 130$ K levels of our HISA features.

The choice of $T_{crit}{}' = 70$ K is empirically based. As noted in Paper I and §2.1, finely-structured H I emission is common in the CGPS data where the total amount of emission is low, i.e., off the plane and at interarm velocities. Without some sort of T_{crit} filtering, the HISA identification software is easily fooled in these regions, flagging many false HISA features adjacent to and between sharp-edged emission. We are confident these are false HISA features, since the absorbing H I would have to be unrealistically cold to absorb against such faint H I backgrounds, and such apparently strong features are far less abundant in brighter emission fields where they should be easier to detect and where more gas should be found generally. There is no single $T_{crit}{}'$ value that excludes all such false HISA while retaining all real HISA. We chose $T_{crit}{}' = 70$ K to balance these two needs, with greater priority placed on the first. For the CGPS, 70 K rejects essentially all false HISA arising from sharp emission edges while keeping most real HISA. The model tests of §3 show that the false HISA rejection is quite successful. A few cases of some real HISA being missed or truncated are discussed below and in Paper II.

2.4.3. Examples

Figures 4-6 illustrate the HISA amplitude extraction process with sample channel maps and spectra that include the same features shown in Figures 1-3. These give the initial H I data, the ΔT amplitudes computed in different stages of the analysis, and the final T_{U}. As the figures show, most of the visually apparent HISA is readily extracted. Some residual HISA remains in T_{U}, but its amplitude is a few K at most; the ΔT value of -40 K extracted for the Perseus Globule is only 2 K weaker than that obtained with the more conservatively-chosen Paper I spatial T_{OFF} boxes (Table 1).

The HISA identification software cannot detect features larger than its CLEAN scales [8 km s^{-1} and 20′] and instead flags only their darkest parts. To overcome this limitation, we feed the T_{U} cubes back into our search algorithms to identify HISA missed on the previous pass (§2.1). Subsequent ΔT extractions are made with unions of HISA flags from all prior search passes but always use the original H I data for T_{OFF}. Three such passes are adequate for the CGPS H I data set. Of the HISA voxels extracted in all three passes, 86.5% were found in the first pass, 11.2% in the second, and only 2.3% in the third.

For brevity, Figures 4-6 display only third-pass results, which differ little from the first pass for this example. A case with more dramatic differences between passes is illustrated in Figures 7 & 8. Features this big require multiple passes to capture. HISA flagging is significantly improved after the first pass in both the spatial and spectral domains. We show only first- and third-pass results here, since the second pass closely resembles the third. After three passes, HISA flagging is incomplete in only a few places due to $T_{U} < 70$ K truncation, mostly near the northern edge of the map (Fig. 7). Aside from minor losses from T_{U} changes, the flagged HISA generally increases, with the fraction of the total flagged per pass being 72.9%, 22.5%, and 4.6% for passes 1, 2, and 3. The smooth $T_{U}(\ell, b, v)$ structure in Figures 7 & 8 shows that our T_{U} estimation method follows the large-scale H I emission brightness reasonably well.

2.4.4. A Note on the Assignment of Structure

We have chosen to attribute fine-scale structure in T_{ON} to ΔT, leaving T_{U} smooth on the scale of the HISA feature. This approach presumes that the absorbing gas is finely structured and the H I background is not, consistent with our adopted HISA identification strategy (§2.1). Such consistency allows the identified HISA structure to be removed so that subsequent search passes do not flag it again. If however some T_{ON} structure arises from T_{U} (e.g., Knee & Brunt 2001), the true ΔT is smoother than we have found. We feel that our

choice of method is reasonable for most circumstances. Small-scale H I emission structure is common in the general ISM but appears minimized in the bright, smooth H I fields where we see most CGPS HISA.

3. Survey Reliability and Completeness

3.1. Motivation

The eye is the first means of identifying HISA features that meet the appropriate criteria, and the search algorithms of §2 were designed primarily to mimic visual detection. However, the eye can be fooled; for example, it often finds false patterns in noise, perhaps due to evolutionary pressures to spot predators (Peebles 1993). We made our HISA search and extraction algorithms as rigorous as possible, but they remain limited by a number of factors, including:

1. $T_{_U}$ **faintness:** To avoid confusion with emission gaps at low column densities, HISA detection is blocked if $T_{_U} < 70$ K. Where this occurs, small features or parts of large ones may be missed.

2. $T_{_U}$ **underestimation:** We assume $T_{_U}$ is not finely structured and estimate it from $T_{_{OFF}}$ voxels surrounding the HISA feature in 3-D. However, many HISA features occur near spectral emission peaks. If $T_{_U} > \langle T_{_{OFF}} \rangle$, then we underestimate $T_{_U}$ and $|\Delta T|$. Both can also be underestimated if HISA flagging is incomplete and an unidentified "penumbra" of faint HISA contaminates $T_{_{OFF}}$.

3. **Noise degradation:** Despite smoothing, some low-amplitude HISA will be lost to noise. Whole features may be missed, or just their cores may be detected, making them appear smaller, clumpier, and more fragmented than they really are. In addition, false HISA detections will be introduced by noise fluctuations at low $|\Delta T|$.

4. **Overlarge Structure:** By design, the HISA search algorithms cannot flag whole features larger than the adopted $20'$ and $8\,\mathrm{km\,s^{-1}}$ CLEAN filter scales. The use of multiple search passes eases this limitation but may not remove it entirely.

5. **Unresolved Structure:** Small-scale HISA structure may be diluted or missed entirely if undersampled. Angular structure down to the $1'$ CGPS beam is seen (Paper I), so smaller-scale structure seems likely. HISA linewidths narrower than the CGPS Nyquist limit of $1.65\,\mathrm{km\,s^{-1}}$ also exist (Knapp 1974; Li & Goldsmith 2003). Such linewidths are rare in random HICA sight lines (e.g., Colgan, Salpeter, & Terzian

1988), but since continuum backgrounds can be brighter than H I backgrounds, HICA can include warmer absorbing gas than HISA, and it's possible that HISA lines may be narrower on average. HISA velocity dilution is thus a real concern in synthesis surveys, while angular dilution will be more severe for single-dish telescopes; no present instrumentation can adequately sample both the angular and velocity structure of HISA.

To evaluate such limitations objectively and quantitatively, we tested our software's ability to extract HISA features from model H I data. Our goal was to understand (1) what fraction of HISA the software detects, (2) how many detections are false positives, and (3) how much the detected features differ in size and amplitude from their input versions.

3.2. Models

The model 21cm spectral line cubes were sums of noisy, positive-amplitude emission backgrounds and noise-free, negative-amplitude absorption features. Gas properties and radiative transfer effects were not considered, as these are irrelevant to the detection software's performance. To test this under varying conditions, 64 randomly-configured model cubes were made. Each cube used the standard CGPS pixel and channel sizes, with dimensions half those of a standard CGPS mosaic cube for computational efficiency: $2.56^\circ \times 2.56^\circ \times 106\,\mathrm{km\,s^{-1}}$ ($512 \times 512 \times 128$ voxels). Sample model data are shown in Figure 9.

3.2.1. Absorption Features

Each model HISA feature was given a cylindrical shape in the H I line cube, with a Gaussian velocity profile and a flat-disk spatial profile convolved with a $60''$ circular beam. Although simple, these angular and velocity profiles are similar enough to typical HISA for testing purposes. The features, known as "hockey pucks" for their usually oblate aspects in the CGPS voxel grid, are parameterized by their unconvolved angular FWHM $\Delta\theta_p$, velocity FWHM Δv_p, and (negative) central amplitude ΔT_p.

2048 hockey pucks were inserted into each model cube with random sizes, amplitudes, and positions. The (ℓ, b, v) and ΔT_p distributions were uniformly random, except that puck overlaps in (ℓ, b, v) were prevented, with a minimum separation of 1 voxel enforced between pucks at an absorption threshhold of 0.005 K. The $\Delta\theta_p$ and Δv_p distributions were skewed toward small features, with relative probabilities of $P(\Delta\theta_p) \propto {\Delta\theta_p}^{-2}$ and $P(\Delta v_p) \propto {\Delta v_p}^{-1}$. This was done to counter the fact that larger pucks have more voxels. We measured angular

and velocity widths locally from each voxel in the performance analysis (§3.3), and $P(\Delta\theta_p)$ and $P(\Delta v_p)$ made our voxel-based size distributions more evenly sampled.

The puck parameter ranges used were $0.1' \leq \Delta\theta_p \leq 60'$, $0.355\,\mathrm{km\,s^{-1}} \leq \Delta v_p \leq 16.0\,\mathrm{km\,s^{-1}}$, and $-1\ \mathrm{K} \geq \Delta T_p \geq -40\ \mathrm{K}$. $\Delta\theta_p = 0.1'$ results in "unresolved" structures that get diluted in the CGPS beam Similarly, $\Delta\theta_p = 0.355\ \mathrm{km\,s^{-1}}$, which would occur for purely thermal H I linewidths at 2.73 K, would be unresolved by the CGPS in velocity. Both cases test the detection limits for fine-scale HISA structure. At the other extreme, the software's sensitivity to structures larger than the $20'$ and $8\,\mathrm{km\,s^{-1}}$ CLEAN filter scales is also tested.

3.2.2. Emission Background

The background emission fields were similarly constructed of random ensembles of cylindrically-symmetric components. These differed from the hockey puck absorption features in that they had positive amplitudes, simple Gaussian angular profiles, and minimum sizes equal to the CLEAN scales. They were also allowed to overlap and fill the entire cube, so we refer to them as emission components rather than discrete features. Size ranges were $20' \leq \Delta\theta_{ec} \leq 120'$ and $8\,\mathrm{km\,s^{-1}} \leq \Delta v_{ec} \leq 20\,\mathrm{km\,s^{-1}}$, with $P(\Delta\theta_{ec})$ and $P(\Delta v_{ec})$ the same as for the HISA pucks. The amplitude range was $+1\ \mathrm{K} \leq \Delta T_{ec} \leq +20\,[V_{ec}/V_{ec,max}]$ K, where the component volume $V_{ec} \equiv {\Delta\theta_{ec}}^2 \Delta v_{ec}$, $V_{ec,max} = (120')^2\,20\,\mathrm{km\,s^{-1}}$, and the ΔT_{ec} distribution was further skewed as $P(\Delta T_{ec}) \propto {\Delta T_{ec}}^{-1}$. These adjustments placed most of the power at large scales, as is seen in real H I emission (Green 1993). 4096 components were summed to make each model cube's emission field. This was subsequently rescaled to give a median brightness temperature of 70 K, so that half the cube on average would allow HISA detections, and $T_{_U} \sim 70$ K effects could be easily studied.

For greater realism, noise was added to the H I model. A 3-D field of uncorrelated Gaussian random voxel noise was convolved with a $60''$ FWHM Gaussian beam and a $1.319\,\mathrm{km\,s^{-1}}$ FWHM Gaussian velocity point spread function (PSF) to mimic the structure of correlated noise in the CGPS data, and the *rms* noise amplitude was scaled to match the 6 K level found in CGPS field centers filled with 100 K emission. Unlike the CGPS noise, the model noise does not vary with distance from field centers, its beam is declination-independent, and its velocity PSF is not the true CGPS velocity PSF, which is the Fourier transform of a Gaussian truncated at 20% of peak amplitude, with an effective FWHM of $1.319\ \mathrm{km\,s^{-1}} = 1.6$ channels. However, none of these differences should seriously affect the performance analysis.

3.3. Analysis

The same procedures used to search for HISA in the CGPS data (§2) were applied to the model data. The software performance was then evaluated by comparing the input and extracted absorption. Each of the 64 model cubes was analyzed separately, and the results were merged afterward to maximize coverage of the model parameter space.

3.3.1. Measurement of HISA Observables

Four observables were extracted from the HISA data: the absorption amplitude ΔT, unabsorbed brightness $T_{_U}$, angular width $\Delta\theta$, and velocity width Δv of the absorption. All four were measured at each voxel (ℓ, b, v) position rather than on a per-feature basis, because the CGPS HISA has complex structure, and properties can vary within one feature. However, $\Delta\theta$ and Δv are still aggregate properties that depend on the local distribution of HISA around them.

The velocity width Δv measures the line FWHM. For each HISA voxel, all HISA contiguous in v at the same (ℓ, b) position is examined to find the channel with maximum $|\Delta T|$. On either side of this channel, the closest channels for which $|\Delta T| \leq 0.5\,|\Delta T|_{max}$ are identified; non-HISA channels with $\Delta T = 0$ are included if necessary. The half-maximum velocities are refined to sub-channel accuracy by linear interpolation. The difference between them is Δv. This Δv is assigned to all HISA voxels in the same velocity grouping at the same (ℓ, b). We make no attempt to correct for instrumental broadening (e.g., Montgomery et al. 1995), since this is nontrivial for the complex line structure of some HISA, and only the narrowest features will be broadened significantly in the CGPS. Figure 10 shows a sample map of Δv. On average, the broader linewidths occur in larger HISA features.

The angular width $\Delta\theta$ measures the diameter of the largest circle containing the (ℓ, b) position and zero non-HISA voxels at the same velocity. This scheme measures the edge-to-edge feature width on a local scale. Unlike the FWHM-based Δv, $\Delta\theta$ uses the full HISA extent. Experiments with an angular FWHM proved too sensitive to complex internal structure in the $\Delta T(\ell, b)$ distribution to be interpreted easily. The resulting $\Delta\theta$ measures can be a bit larger than the FWHM-based hockey puck $\Delta\theta_p$, especially for large puck amplitudes ΔT_p, but since the same $\Delta\theta$ measure is taken of the HISA model inputs and outputs, the method is internally consistent.

Figure 10 illustrates how $\Delta\theta$ is measured. From each HISA voxel, the angular offset $\theta_{\rm off}$ to the nearest non-HISA voxel with the same velocity is found. $\theta_{\rm off}(\ell, b)$ maps HISA "skeletons" whose ridge-like maxima equal half the local width of the feature. To build a $\Delta\theta$

map, we step over all (ℓ, b) positions and write $2\,\theta_{\rm off}(\ell, b)$ to all points (ℓ', b') in a new map for which $\sqrt{(\ell' - \ell)^2 + (b' - b)^2} \leq \theta_{\rm off}(\ell, b)$, where the largest imposed $2\,\theta_{\rm off}$ value is always retained. This yields the angular width of HISA filamentary structure at that velocity.

3.3.2. Performance Measures

The HISA extraction software's performance was measured in three ways: the "throughput" fraction f_{det} of model HISA detected; the "true fraction" f_{true} of HISA detections corresponding with model input; and the "drifts" $\Delta\Delta T$, $\Delta T_{_U}$, $\Delta\Delta\theta$, and $\Delta\Delta v$ between input and output properties. All were measured as functions of the HISA observables $(\Delta T, T_{_U}, \Delta\theta, \Delta v)$, which define a 4-D parameter space in which the software performance is evaluated.

The (ℓ, b, v) positions and $(\Delta T, T_{_U}, \Delta\theta, \Delta v)$ properties were first tabulated for all HISA voxels in both the input and output model (ℓ, b, v) cubes. The 4-D voxel count histograms $N_{in,all}(\Delta T_{in}, T_{_U in}, \Delta\theta_{in}, \Delta v_{in})$ and $N_{out,all}(\Delta T_{out}, T_{_U out}, \Delta\theta_{out}, \Delta v_{out})$ were constructed from these voxel tables, using bin dimensions of $2.5\ {\rm K} \times 2.5\ {\rm K} \times 0.5' \times 0.5\,{\rm km\,s^{-1}}$. In parallel, the voxel count histogram $N_{in,det}$ was made from all input voxels with output at the same (ℓ, b, v), and $N_{out,true}$ was made from all output voxels with input at the same (ℓ, b, v). Throughputs and true fractions were then derived as $f_{det} \equiv N_{in,det}/N_{in,all}$ and $f_{true} \equiv N_{out,true}/N_{out,all}$. From the subset of voxels appearing in both the input and output HISA cubes, four 4-D drift histograms of the average changes undergone by ΔT_{in}, $T_{_U in}$, $\Delta\theta_{in}$, and Δv_{in} as functions of $(\Delta T_{in}, T_{_U in}, \Delta\theta_{in}, \Delta v_{in})$ were assembled, e.g., as $\Delta\Delta T \equiv \langle \Delta T_{out} - \Delta T_{in} \rangle$, with the average taken over all HISA voxels in (ℓ, b, v) with the same $(\Delta T_{in}, T_{_U in}, \Delta\theta_{in}, \Delta v_{in})$ properties.

The 4-D performance histograms were computed for all 64 HISA models, which were identical apart from different random number inputs. The results were merged together into a single set of histograms and smoothed with 4-D Gaussians to improve the performance measure reliability and coverage of the parameter space. Variable smoothing scales were used, because the parameter space coverage was sparser in some areas than in others. The smoothing FWHM were $0.5|\Delta T|$, $0.5|T_{_U} - 70\ {\rm K}|$, $0.5\Delta\theta$, and $0.5\Delta v$, with minimum values of 6.0 K, 6.0 K, 1.0′, and $1.319\,{\rm km\,s^{-1}}$ to match the model T_{rms} and CGPS resolution. This scheme preserved structure in the well-sampled parts of the parameter space and interpolated it smoothly elsewhere.

$N_{out,all}(\Delta T_{out}, T_{_U out}, \Delta\theta_{out}, \Delta v_{out})$ histograms were also computed for all 36 mosaic cubes of real CGPS HISA and summed together to assess the distribution of observed HISA in the survey. As with the HISA feature catalog in Paper II, sight lines with $T_{_C} > 20$ K were excluded.

3.4. Results

3.4.1. Model Parameter Distributions

To examine general $(\Delta T, T_{_U}, \Delta\theta, \Delta v)$ parameter distributions and trends, we made 2-D projections of the unsmoothed model $N_{in,all}$ and $N_{out,all}$ and the real CGPS $N_{out,all}$ by summing the counts along the 2 other axes in the 4-D parameter space. A number of these 2-D projections are shown in Figure 11.

The input models fill ranges of $0 > \Delta T > -40$ K, $30 < T_{_U} < 150$ K, $0 < \Delta\theta < 61'$, and $0.8 < \Delta v < 16\,\mathrm{km\,s^{-1}}$, with peaks at $T_{_U} = 70$ K and small $|\Delta T|$ and $\Delta\theta$; the Δv distribution was relatively flat. The peaked ΔT and $\Delta\theta$ distributions occurred despite the shapes chosen for the puck property distributions (§3.2). The low-$|\Delta T|$ peak is due to faint HISA in feature line wings and spatial envelopes. The low-$\Delta\theta$ peak results from $\Delta\theta$ being measured from HISA voxels above a minimum $|\Delta T|$, which makes pucks appear smaller off the line center. In a similar way, pucks with the same $\Delta\theta_p$ have greater $\Delta\theta$ if $|\Delta T_p|$ is larger, and no voxels with $\Delta\theta < 1.5'$ and $\Delta T < -20$ K are found.

The extracted model ΔT peak is shifted to ~ -10 K. A tail of strong absorption extends to $\Delta T \sim -60$ K. Although they account for only 2% of the total HISA voxels, these $\Delta T < -40$ K points demonstrate that some ΔT drift occurs. The extracted $T_{_U}$ is truncated at $\lesssim 70$ K but otherwise appears unchanged from the input model. Large $\Delta\theta$ and Δv values are both truncated as predicted in §3.1, although less severely for $\Delta\theta$, since the non-Gaussian angular profiles better survive the CLEAN process. Iterative extraction (§2) allows much of the puck structure $> 20'$ to be recovered here, but the purely Gaussian puck velocity profiles with $\Delta v_p > 8\,\mathrm{km\,s^{-1}}$ are CLEANed out of the data with great efficiency. $\Delta\theta$ peaks at the same location as the input data but is more concentrated. Δv is also now concentrated toward low values.

3.4.2. Real Parameter Distributions

The CGPS HISA ΔT has a larger range than in the extracted models, due to a few very strong features like GHISA 079.88+0.62+02 and GHISA 091.90+3.27−03 (see Paper II for feature details). Its ΔT peak is similar to the models'. The CGPS HISA is truncated for $T_{_U} < 70$ K as well as $T_{_U} > 135$ K, where the maximum H I brightness is reached. $\Delta\theta$ peaks at the same scale as the model output but has a lower maximum scale, perhaps because real HISA is more porous. The model output Δv range is slightly exceeded. The Δv peak and maximum value are both a little broader than for the extracted model HISA.

The CGPS HISA fills almost the same parameter space as the extracted model HISA. Since some model parameter ranges are larger on input than output, the ranges of real HISA properties may exceed those observed in the CGPS. HISA with $T_U < 70$ K is already known (e.g., Knee & Brunt 2001), and HISA with $\Delta v > 8\,\mathrm{km\,s^{-1}}$, $\Delta\theta > 33'$, or $|\Delta T| > 80$ K is possible, although $|\Delta T| \leq p\,T_U + T_C - T_S$ is required by the 4-component radiative transfer equation (Paper II, Eqn. 1), where T_S is the spin (excitation) temperature of the absorbing gas. Also, $\Delta\theta$ is limited by feature porosity. GHISA 091.90+3.27−03 exceeds the $\Delta\theta$ limit in gross extent but is not completely solid. Larger features are known (e.g., the Riegel & Crutcher 1972 "cold cloud" toward the Galactic center), but their porosity at $1'$ resolution has not been reported.

The input models had no built-in correlations of feature properties, and the same is largely true for the real HISA. Certainly ΔT and T_U are not related in the CGPS, except that the strongest ΔT's prefer some T_U values over others. However, the peak CGPS $\Delta\theta$ and Δv both increase gradually with $|\Delta T|$ out to $\Delta T = -40$ K, $\Delta\theta = 20'$, and $\Delta v = 4\,\mathrm{km\,s^{-1}}$. These trends have considerable scatter, and there are weaker versions in the extracted model HISA. But if they reflect real HISA behavior, then stronger absorption is more likely to have larger contiguous angular structure or broader linewidths, although $\Delta\theta$ and Δv do not correlate as well with each other as they do with ΔT.

3.4.3. Throughput and True Fraction

Figure 12 presents selected 2-D slices through the 4-D parameter space to illustrate the behavior of the throughput f_{det} and true fraction f_{true}. These have been smoothed as described in §3.3.2. We find that most HISA is detected if it is significantly stronger than the noise, larger than a few beams, narrower than a few $\mathrm{km\,s^{-1}}$, and has $T_U \gtrsim 80$ K. Furthermore, the vast majority of detected HISA is reliable, except for HISA that can be mimicked by beam-scale noise fluctuations.

The throughput is high for much of the parameter space: $f_{det} \gtrsim 0.80$ where $\Delta T \lesssim -20$ K, $T_U \gtrsim 80$ K, $\Delta\theta \gtrsim 5'$, and $\Delta v \lesssim 3.5\,\mathrm{km\,s^{-1}}$, reaching a maximum of ~ 0.99 where all of these criteria are well-met. Where one or more of them is not met, f_{det} drops rapidly, with $f_{det} \to 0$ for $\Delta T \gtrsim -2$ K, $T_U \lesssim 60$ K, $\Delta\theta \lesssim 1'$, or $\Delta v \gtrsim 8\,\mathrm{km\,s^{-1}}$. Most of this behavior can be explained as losing features in the noise, underbright T_U, or overbroad linewidths poorly fitted by the spectral search method's W_{broad} $[4\,\mathrm{km\,s^{-1}}]$ Gaussians (§2.2). However, low f_{det} seems to occur for low $\Delta\theta$ even when $|\Delta T|$ is large. This suggests some beam-scale HISA features may be missed by our search if they are isolated from larger structures. The Perseus HISA Globule of Paper I is detected easily, but it is also attached to the complex

GHISA 139.01+0.96−40.

By contrast, narrow-line HISA is detected with great efficiency: high f_{det} is found for Δv as low as the $0.8\,\mathrm{km\,s^{-1}}$ CGPS channel width, so long as $\Delta T \lesssim -10$ K. To allay the concerns of Li & Goldsmith (2003), lines narrower than this should also be detectable if their intrinsic amplitudes are larger to compensate for spectral dilution. A thermally broadened $T_{_S} = 10$ K HISA line would be diluted by a factor of 2.4 but easily detected if $\Delta T \lesssim -24$ K. HISA of this strength or greater is common in the CGPS.

The true fraction has much simpler behavior: $f_{true} \rightarrow 1.0$ almost everywhere that $\Delta T \lesssim -20$ K, $\Delta\theta \gtrsim 3'$, or $\Delta v \gtrsim 3\,\mathrm{km\,s^{-1}}$. If none of these holds, noise fluctuations in the data produce significant false positive detections, with $f_{true} \gtrsim 0.2$ in the worst cases.

3.4.4. Parameter Drift

Figure 13 illustrates trends in the parameter drifts $\Delta\Delta T$, $\Delta\Delta\theta$, and $\Delta\Delta v$. Since $T_{_U} = T_{_{ON}} - \Delta T$ (§2.4) and $T_{_{ON}}$ is fixed, $\Delta T_{_U} = -\Delta\Delta T$. With minor exceptions, the behavior of all the parameter drifts is fairly simple: $|\Delta T|$ and $T_{_U}$ are often underestimated by a few K in well-detected features, while incomplete detections of large features (e.g., Fig. 9) cause $\Delta\theta$ and Δv to be underestimated as well.

The drift in ΔT is positive, i.e., toward reduced amplitudes, if these three conditions are met: $|\Delta T| \gtrsim$ the 6 K noise level, $T_{_U} \gtrsim 80$ K, and $\Delta\theta \gtrsim 2'$. If one of them is not met, $\Delta\Delta T < 0$. There is no strong dependence on Δv. The amount of drift is typically a few K, with a range of ± 10 K in most areas but more negative for $T_{_U} < 65$ K. The $\Delta\Delta T$ behavior has a similar shape to f_{det} above, suggesting that detection sensitivity governs ΔT drift. Features with intrinsically low $|\Delta T|$, low $T_{_U}$, or very small $\Delta\theta$ appear to have larger $|\Delta T|$ (and $T_{_U}$) if they are detected. But f_{det} shows most are *not* detected; those that are represent a biased sample in which $|\Delta T|$ and $T_{_U}$ happened to be boosted in the right direction to make them detectable. By contrast, easily detectable features appear to have lower $|\Delta T|$ than they should. Since $\Delta T_{_U} = -\Delta\Delta T$, their $T_{_U}$ is underestimated as well; one of the mechanisms noted in §3.1 may be to blame. But whether positive or negative, the magnitude of $\Delta\Delta T$ is usually only a few K. The large drifts that produced the ΔT outliers in the model $N_{out,all}$ results (§3.4.1; Fig. 11) are exceptional cases.

The drift in $\Delta\theta$ is negative everywhere that $\Delta\theta > 1'$. It covers a range of $-50' \lesssim \Delta\Delta\theta \lesssim +1'$, becoming steadily more negative for larger $\Delta\theta$, with only minor dependencies on other parameters. The small positive drifts occur when $\Delta\theta < 1'$ HISA is augmented by beam-scale noise fluctuations; $\Delta\theta < 1'$ can occur in the line wings of $1'$ features, since pucks

appear smaller off the line center (§3.4.1). The much larger negative drifts are from those large features that are incompletely detected, as confirmed by visual inspection of the (ℓ, b, v) data (e.g., Fig. 9). Some of these partial detections are caused by $T_U \sim 70$ K boundaries, but most are from noise. 2- and 3-σ noise fluctuations frequently poke holes in fairly strong features, reducing their $\Delta\theta$ measures. This is especially common in the line wings where $|\Delta T|$ is less.

The drift in Δv is positive only where both $\Delta\theta < 2'$ and $\Delta v < 1\,\mathrm{km\,s^{-1}}$ and is negative everywhere else. It covers a range of $-13\,\mathrm{km\,s^{-1}} < \Delta v < +0.2\,\mathrm{km\,s^{-1}}$, becoming steadily more negative for larger Δv, with only minor dependencies on other parameters. As with $\Delta\Delta\theta$, noise degradation is a major cause of this $\Delta\Delta v$ trend. But in addition, the spectral HISA search itself is optimized for the detection of HISA linewidths $\lesssim 4\,\mathrm{km\,s^{-1}}$, and Gaussian features broader than $8\,\mathrm{km\,s^{-1}}$ are CLEANed out almost entirely (§3.4.3). Lastly, features near emission peaks may have T_U underestimated (§3.1), leading to $|\Delta T|$ underestimates in line wings, and thus Δv underestimation. Since our T_U estimation is not a simple linear interpolation (§2.4), this effect may not be as severe as that noted by Levinson & Brown (1980), but the $\Delta\Delta T > 0$ results above for well-detected features suggest it is not zero either.

Levinson & Brown (1980) also note that T_U gradients will cause Δv to be narrower than the HISA optical depth profile FWHM, and the line center in ΔT will appear shifted from τ_{max}, the maximum optical depth. However, the performance of our HISA software is only concerned with ΔT, so these biases do not apply here. And while our voxel-based evaluation method is not able to track changes of position, visual inspection shows that filaments and other structures within features do not shift between input and output; only the centroids of whole features may shift if the features are not completely detected.

3.5. Reliability and Completeness

These results have many uses. In addition to statistically describing how well HISA is detected, they can be applied directly to particular features to assess their reliability and completeness. The former is done by measuring $(\Delta T, T_U, \Delta\theta, \Delta v)$ at each HISA voxel (ℓ, b, v) position and interpolating f_{true} from the 4-D histogram described in §3.4.3. f_{true} is the likelihood that a HISA detection represents real absorption. We have determined f_{true} for each CGPS HISA voxel. Figure 14 shows f_{true} contours on a sample HISA feature. The HISA detection reliability in this case is quite high. In Paper II, we use $f_{true}(\ell, b, v)$ in analyses of total CGPS HISA coverage and the distributions of weak and strong absorption. We also assess the completeness of our HISA detections. The actual detection fraction f_{det} is not recoverable from observed data, but we consider the fraction of detections with $\langle T_U \rangle > 80$ K,

since such HISA has $f_{det} > 0.8$ if its size and strength are appreciable (§3.4.3).

4. Conclusions

We have described algorithms that identify and extract H I self-absorption (HISA) features in high-resolution H I 21cm line data cubes. These algorithms were designed to carry out a HISA survey of cold H I in the initial $73^\circ \times 9^\circ$ phase of the arcminute-resolution Canadian Galactic Plane Survey (CGPS), but they should have more general applicability.

Our search algorithms use CLEAN-based spatial and spectral filtering to remove large-scale emission structure and identify HISA as significant negative residuals. Features identified in both spectral and spatial domains are flagged as HISA, and the unabsorbed brightness $T_{_U}$ along the feature sightline is estimated from a 3-D interpolation of the OFF-feature brightness temperature $T_{_{OFF}}$. HISA detections in overly noisy regions are rejected, as are those for which $T_{_U} < 70$ K, lest significant false detections result from gaps between sharply-structured emission features with faint backgrounds. In order to capture features larger than the CLEAN filter scale, identified HISA is removed and the search process is repeated; a total of three such passes suffices for the CGPS data.

We performed detailed tests of our HISA-finding software with model data to determine its detection limits, false positive rates, and measurement biases as functions of feature size, amplitude, and background field brightness. The tests show that HISA is well detected within the software design criteria, with high detection rates for HISA significantly stronger than the noise level, larger than a few beams, narrower than a few $\mathrm{km\,s^{-1}}$, and with $T_{_U} \gtrsim 80$ K. At the same time, the bulk of HISA detections are reliable, with very low false positive rates in most parts of the parameter space except those occupied by beam-scale noise fluctuations. Measurement drifts are small in well-detected features, with $T_{_U}$ underestimated by a few K due to contamination of $T_{_{OFF}}$ by faint, undetected HISA near the feature. Where detections are truncated by noise fluctuations or faint $T_{_U}$, the bias may be somewhat larger. Incomplete detections also make features appear smaller in angular size and linewidth than in reality due to truncation.

This paper is the third in an ongoing series investigating HISA at high resolution in the Galactic plane. A companion paper (Paper II) presents HISA survey results for the CGPS. Subsequent papers will further analyze the CGPS HISA and also examine HISA in CGPS extensions and in the VLA Galactic Plane Survey.

We thank W. McCutcheon, T. Landecker, and J. Stil for a number of useful discussions

on this project, and the anonymous referee for constructive comments on the manuscript. J. Stil helped with multipanel figure layout. We are very grateful to R. Gooch for tireless computing support, including continued expansion of the capabilities of the Karma visualization software package (Gooch 1996)[1], which was used extensively for this work. The Dominion Radio Astrophysical Observatory is operated as a national facility by the National Research Council of Canada. The Canadian Galactic Plane Survey (CGPS) is a Canadian project with international partners. The CGPS is described in Taylor et al. (2003)[2]. The main CGPS data set is available at the Canadian Astronomy Data Centre[3]. The CGPS is supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

A. Corrections to Paper I Results

The H I data in the vicinity of ($\ell = 140°, b = +1°$) have been revised from those used in Paper I. A single synthesis field was assigned the wrong flux scale in the H I data used in that paper, and this error was not discovered until after publication. As a result, the HISA amplitudes presented in Paper I for the Perseus HISA Complex and Globule features were in error, and the correct HISA amplitudes are smaller than those found in Paper I. With revised data, these features have warmer spin temperatures and lower optical depths than those derived in Paper I, but the column densities and masses are only mildly affected. Table 1 lists the corrected results for both features. The correct Globule spectrum is plotted in Figure 1, and the positions of both features are marked in Figure 5. The Local HISA Filament presented in Paper I was unaffected by this problem.

REFERENCES

Baker, P. L., & Burton, W. B. 1979, A&AS, 35, 129

Bania, T. M., & Lockman, F. J. 1984, ApJS, 54, 513

Colgan, S. W. J., Salpeter, E. E., & Terzian, Y. 1988, ApJ, 328, 275

Feldt, C. 1993, A&A, 276, 531

[1]See also http://www.atnf.csiro.au/karma.

[2]Additional information is available online at http://www.ras.ucalgary.ca/CGPS.

[3]See http://cadcwww.hia.nrc.ca/cgps.

Gibson, S. J. 2002, ASP Conf. Ser. 276, *Seeing Through the Dust: the Detection of H I and the Exploration of the ISM in Galaxies*, eds. A. R. Taylor, T. L. Landecker, & A. G. Willis, 235

Gibson, S. J., Taylor, A. R., Dewdney, P. E., & Higgs, L. A. 2000, ApJ, 540, 851 (Paper I)

Gibson, S. J., Taylor, A. R., Higgs, L. A., Brunt, C. M., & Dewdney, P. E. 2005, ApJ, submitted (Paper II)

Gibson, S. J., Taylor, A. R., Stil, J. M., Higgs, L. A., Dewdney, P. E., & Brunt, C. M. 2004, in *How Does the Galaxy Work? A Galactic Tertulia with Don Cox and Ron Reynolds*, eds. E. J. Alfaro, E. Pérez, & J. Franco (Dordrecht: Kluwer Academic Publishers), 47

Gooch, R. 1996, ASP Conf. Ser. 101, *Astronomical Data Analysis Software and Systems V*, eds. G. H. Jacoby & J. Barnes, 80

Green, D. A. 1993, MNRAS, 262, 327

Hasegawa, T., Sato, F., & Fukui, Y. 1983, AJ, 88, 658

Högbom, J. A. 1974, A&AS, 15, 417

Kavars, D. W., Dickey, J. M., McClure-Griffiths, N. M., Gaensler, B. M., & Green, A. J. 2003, ApJ, 598, 1048

Kerton, C. R. 2005, ApJ, accepted

Knapp, G. R. 1974, AJ, 79, 527

Knee, L. B. G., & Brunt, C. M. 2001, Nature, 412, 308

Levinson, F. H., & Brown, R. L. 1980, ApJ, 242, 416

Li, D., & Goldsmith, P. F. 2003, ApJ, 585, 823

McClure-Griffiths, N. M., Green, A. J., Dickey, J. M., Gaensler, B. M., Haynes, R. F., & Wieringa, M. H. 2001, ApJ, 551, 394.

McCutcheon, W. H., Shuter, W. L. H., & Booth, R. S. 1978, MNRAS, 185, 755

Minter, A. H., Lockman, F. J., Langston, G. I., & Lockman, J. A. 2001, ApJ, 555, 868

Montgomery, A. S., Bates, B., & Davies, R. D. 1995, MNRAS, 273, 449

Peebles, P. J. E. 1993, *Principles of Physical Cosmology* (Princeton: Princeton University Press), 35

Peters, W. L., & Bash, F. N. 1987, ApJ, 317, 646

Riegel, K. W., & Crutcher, R. M. 1972, A&A, 18, 55

Steer, D. G., Dewdney, P. E., & Ito, M. R. 1984, A&A, 137, 159

Taylor, A. R., et al. 2003, AJ, 125, 3145

Taylor, A. R., Stil, J. M., Dickey, J. M., McClure-Griffiths, N. M., Martin, P. G., Rothwell, T., & Lockman, F. J. 2002, ASP Conf. Ser. 276, *Seeing Through the Dust: the Detection of H I and the Exploration of the ISM in Galaxies*, eds. A. R. Taylor, T. L. Landecker, & A. G. Willis, 68

van der Werf, P. P., Goss, W. M., & Vanden Bout, P. A. 1988, A&A, 201, 311

This preprint was prepared with the AAS LaTeX macros v5.2.

Table 1. Corrected Perseus HISA Complex and Globule Properties*

	Perseus Complex		Perseus Globule	
H I Data:	Paper I	Revised	Paper I	Revised
Input Parameters†				
T_{ON} [K]	69	71	47	62
T_{OFF} [K]	107	99	112	104
ΔT [K]	-38	-28	-65	-42
Derived Gas Properties ($p = f_n = 1$)				
T_S [K]	$45 - 61$	$49 - 65$	$32 - 35$	$41 - 43$
τ	$0.83 - 1.37$	$0.71 - 1.22$	$1.43 - 1.54$	$0.94 - 0.99$
N_{HISA} [10^{20} cm^{-2}]	$3.2 - 7.3$	$3.0 - 6.8$	$2.2 - 2.6$	$1.8 - 2.0$
n_{HISA} [cm^{-3}]	$89 - 65$	$81 - 62$	$124 - 115$	$99 - 94$
M_{HISA} [M$_\odot$]	$31 - 111$	$32 - 106$	$0.60 - 1.09$	$0.53 - 0.80$
Derived Gas Properties ($f_n = 0.01$, Maximum Total Mass)				
T_S [K]	2.7	2.7	2.7	2.7
τ	7.0	6.9	2.5	2.4
N_{HISA} [10^{20} cm^{-2}]	1.7	1.6	0.33	0.32
n_{HISA} [cm^{-3}]	15	15	15	15
M_{HISA} [M$_\odot$]	26	25	0.14	0.12
N_{tot} [10^{20} cm^{-2}]	170	160	33	32
n_{tot} [cm^{-3}]	1500	1500	1500	1500
M_{tot} [M$_\odot$]	5200	5000	28	25

*This table follows the format of Table 1 in Paper I.

†Only input parameters that have changed from Paper I are shown.

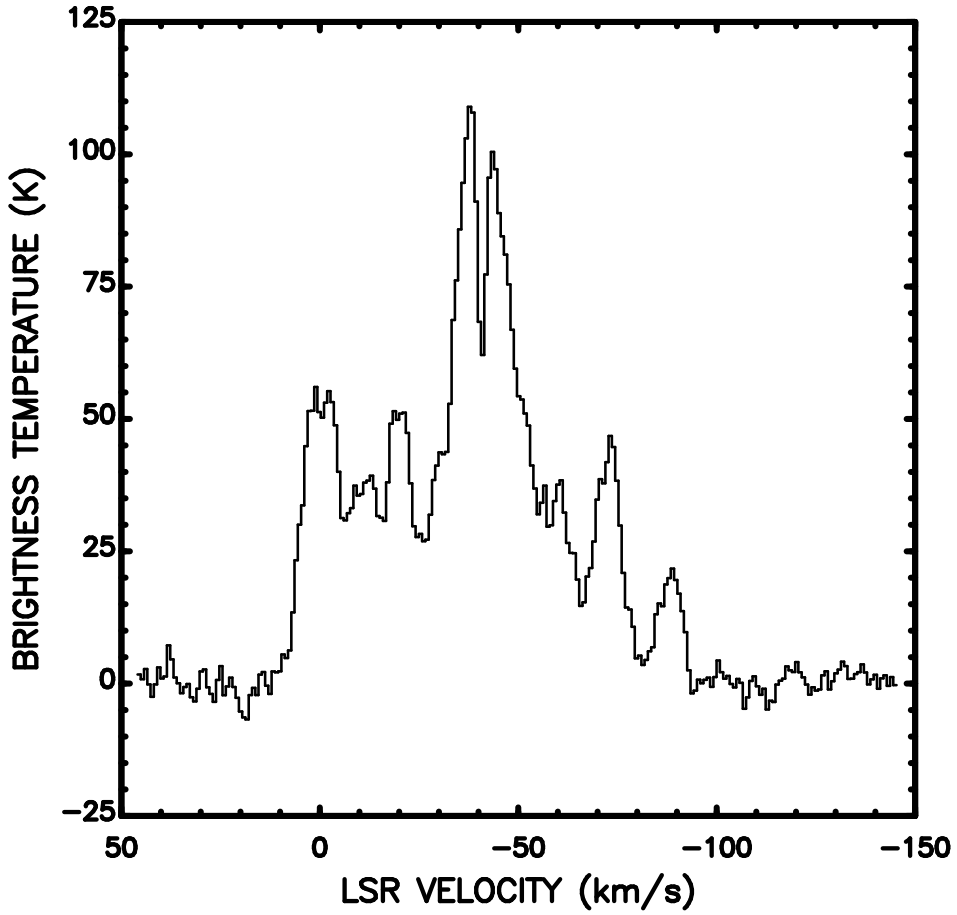

Fig. 1.— The full-resolution observed spectrum $O(k)$ at the ($\ell = 139.635°$, $b = 1.185°$) position of the Perseus HISA Globule of Paper I. The feature's absorption amplitude has changed from Paper I due to correction of a data processing error that had no serious impact on the derived results (see Appendix A).

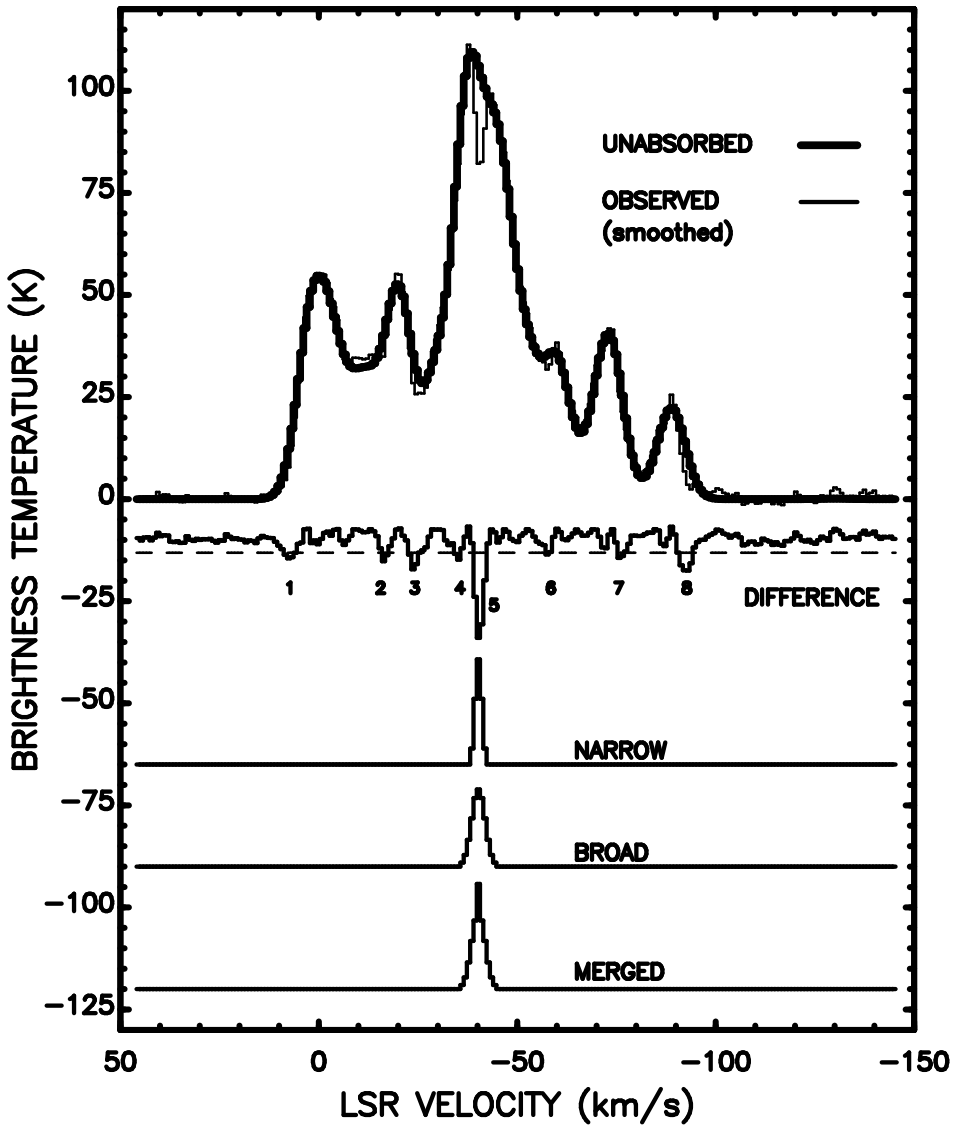

Fig. 2.— Velocity profiles showing HISA spectral detection stages for the Perseus HISA Globule position. The derived unabsorbed spectrum $U(k)$ and the spatially smoothed observed spectrum $S(k)$ are shown in the top portion of the figure. Below them is the residual or difference spectrum $R(k)$ (zero level at −10 K). The dashed line gives the level below which HISA is suspected. Eight channel segments are indicated where this is the case. Gaussian fitting was accepted only for segment 5 (see text), and the resulting *narrow* and *broad* HISA spectra are shown below this (zero levels at −60 K and −90 K, respectively). Finally, the "detected" HISA spectrum is shown at bottom (zero level at −120 K).

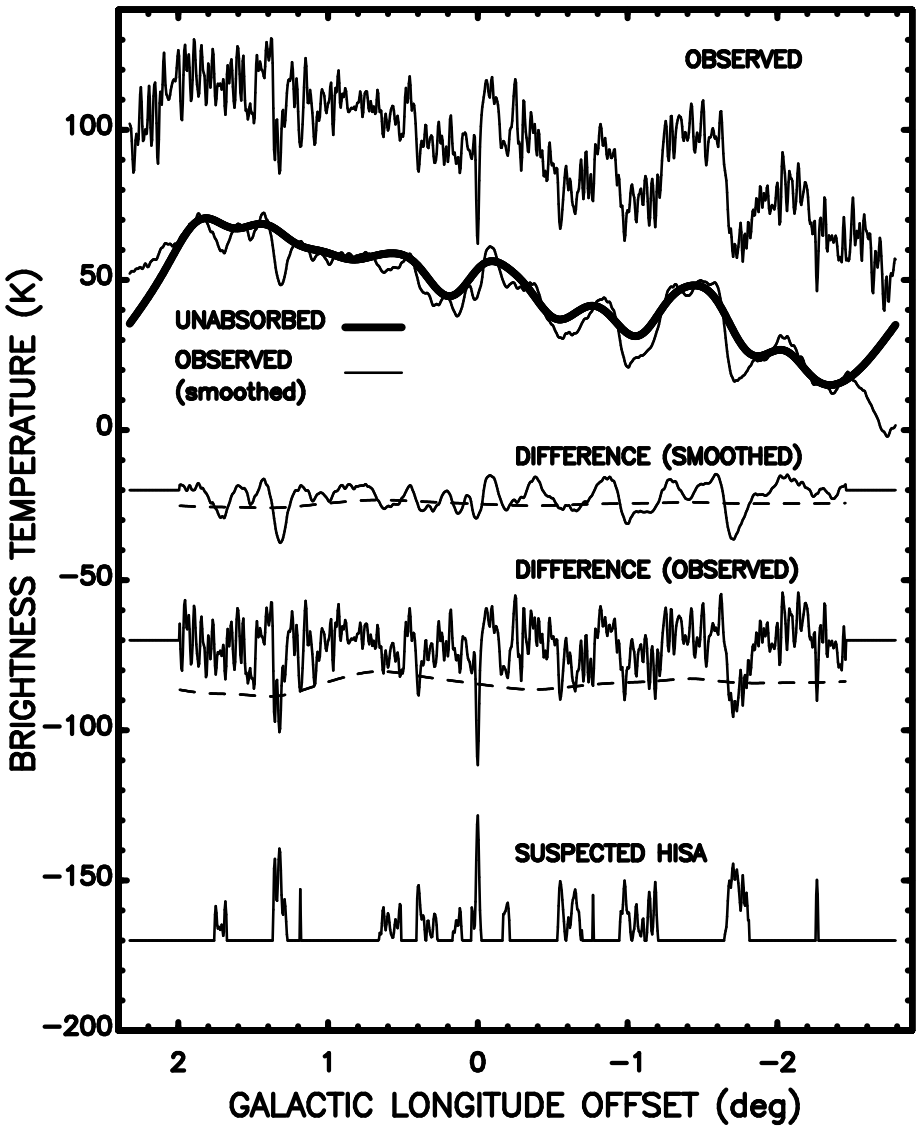

Fig. 3.— Latitude profiles showing HISA spatial detection stages for the Perseus HISA Globule position (at which the Galactic latitude offset = 0°). From top to bottom, cuts are taken through the observed channel map $O(i,j)$, the smoothed observed map $S(i,j)$ and derived unabsorbed map $U(i,j)$ (zero levels at −50 K), the smooth difference map $S(i,j)-U(i,j)$ (solid) and noise truncation level $-2\sigma_{sm}$ (dashed) (zero level at −20 K), the unsmooth difference map $O(i,j)-U(i,j)$ (solid) and noise truncation level $-2\sigma_{obs}$ (dashed) (zero level at −70 K), and the suspected HISA features map prior to final amplitude culling (zero level at −170 K). Fast Fourier transforms (FFTs) used in the CLEANing process make border areas of the map unusable after $S(i,j)$ and $U(i,j)$ are determined.

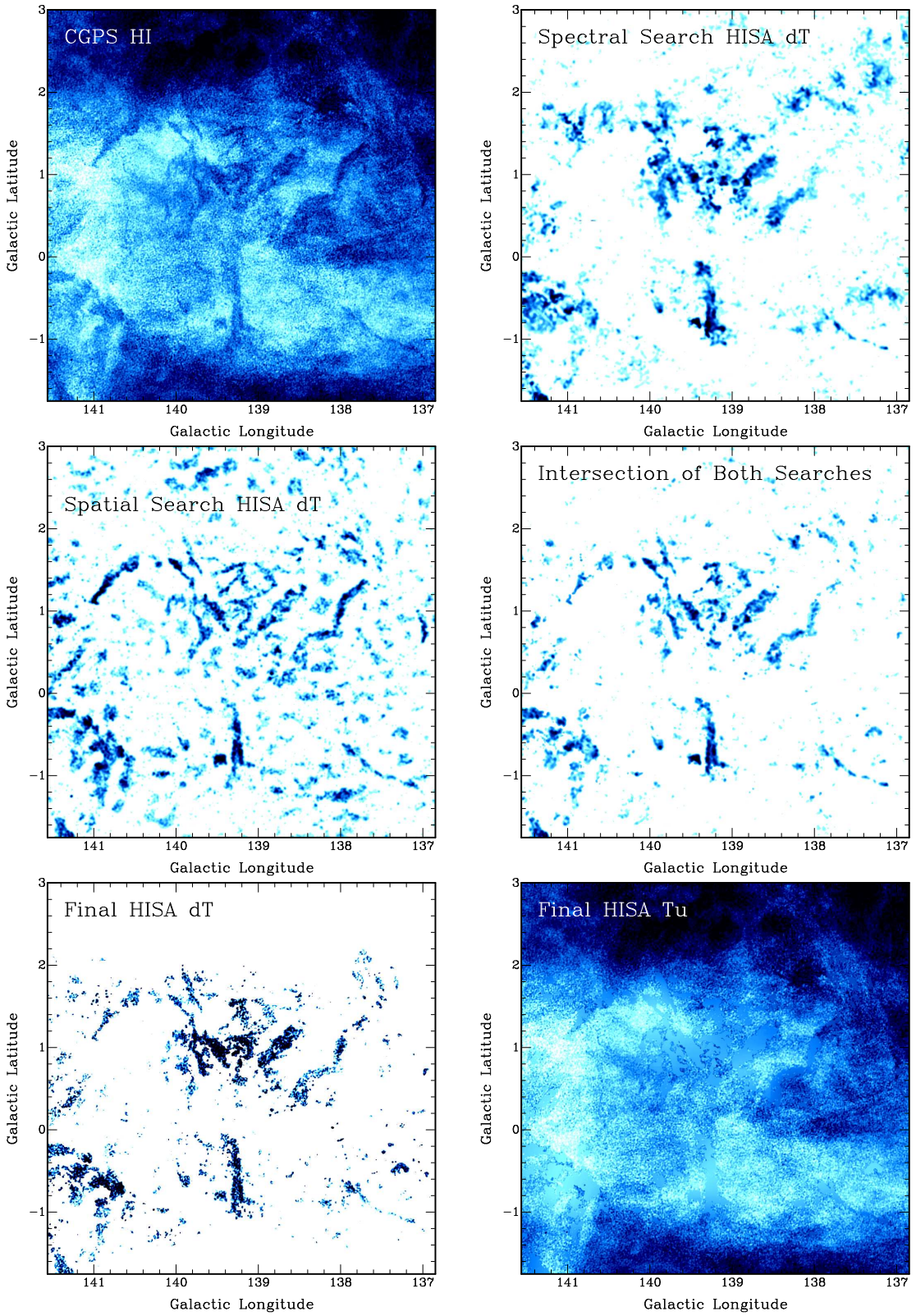

Fig. 4.— (ℓ, b) channel maps of sample Perseus HISA at -41 km s^{-1}, showing the same area as Figure 1 of Paper I. The panels give CGPS H I, HISA ΔT from the spectral search (§2.2), spatial search (§2.3), and their intersection, and the final ΔT and T_U from the full 3-D extraction (§2.4). Intensity ranges are +40 to +130 K for the first and last panels and −20 to 0 K for the rest, from black to white in all cases.

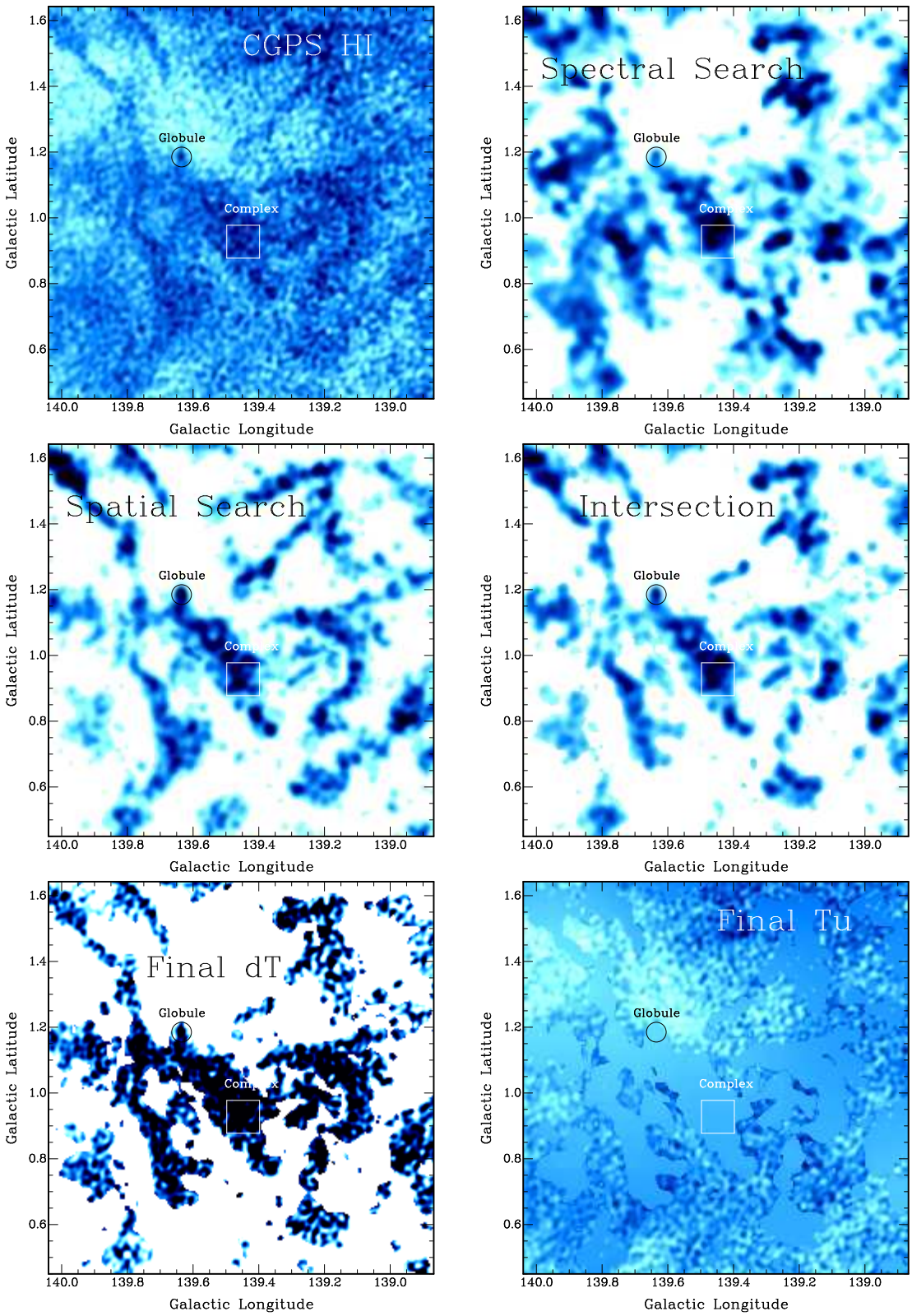

Fig. 5.— Detailed views of each panel in Figure 4, showing the same area as Figure 2 of Paper I. The Perseus HISA Complex and Globule positions of that paper are marked. Sample velocity spectra of the Globule are given in Figure 6.

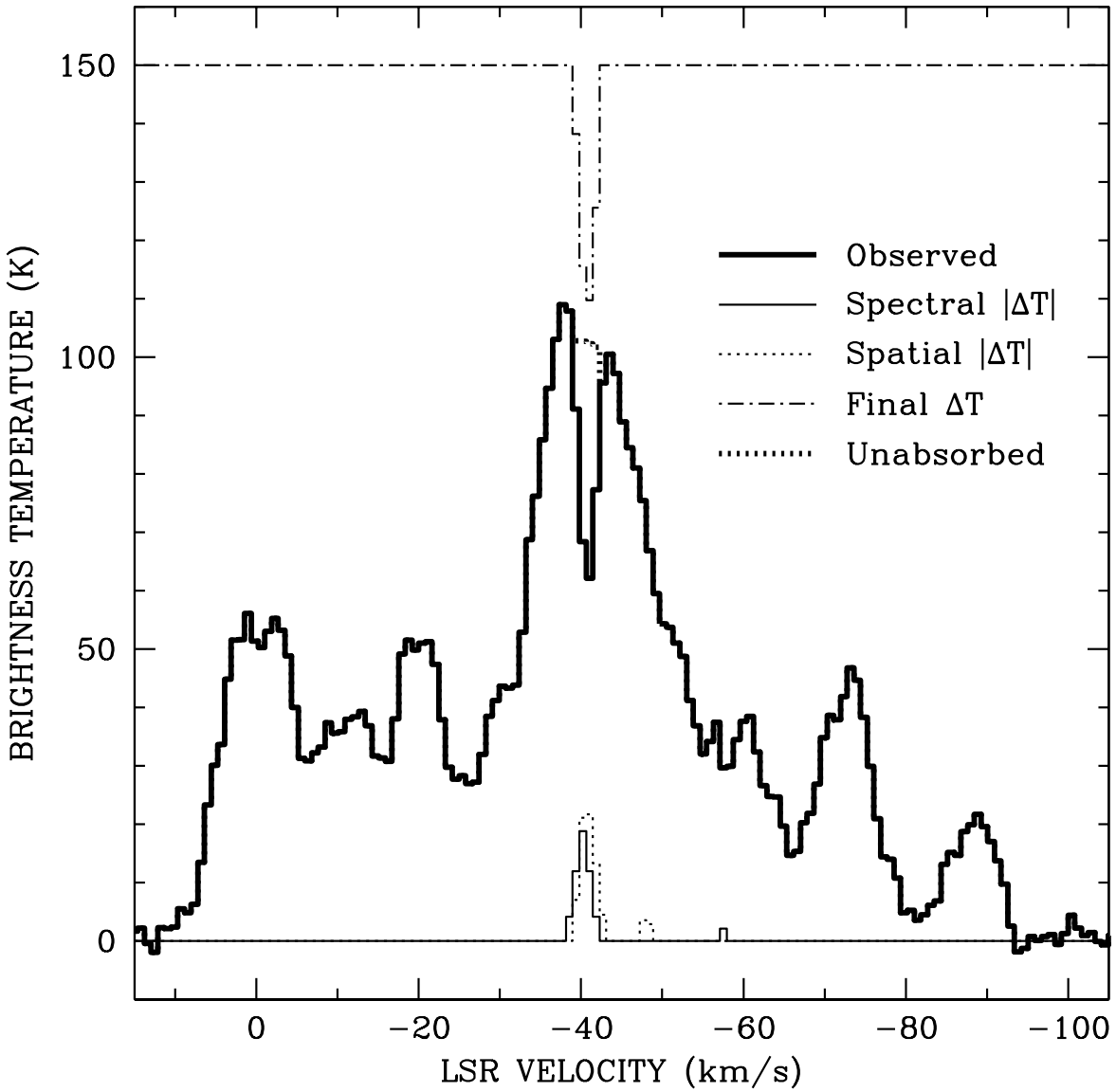

Fig. 6.— Single-pixel velocity spectra at the Perseus HISA Globule position (ℓ = 139.635°, b = 1.185°), showing CGPS H I, HISA $|\Delta T|$ from the spectral (§2.2) and spatial searches (§2.3), and the final ΔT and T_U from the full 3-D extraction (§2.4); the final ΔT zero point has been shifted to 150 K for clarity. Although T_U appears below the observed T_{OFF} at the HISA velocity edges, this spectrum shows only a small subset of the T_{OFF} voxels that surround the feature in 3-D. T_U is estimated from the entire 3-D T_{OFF} set, a larger sample of which is shown in the corresponding channel map in Figure 5.

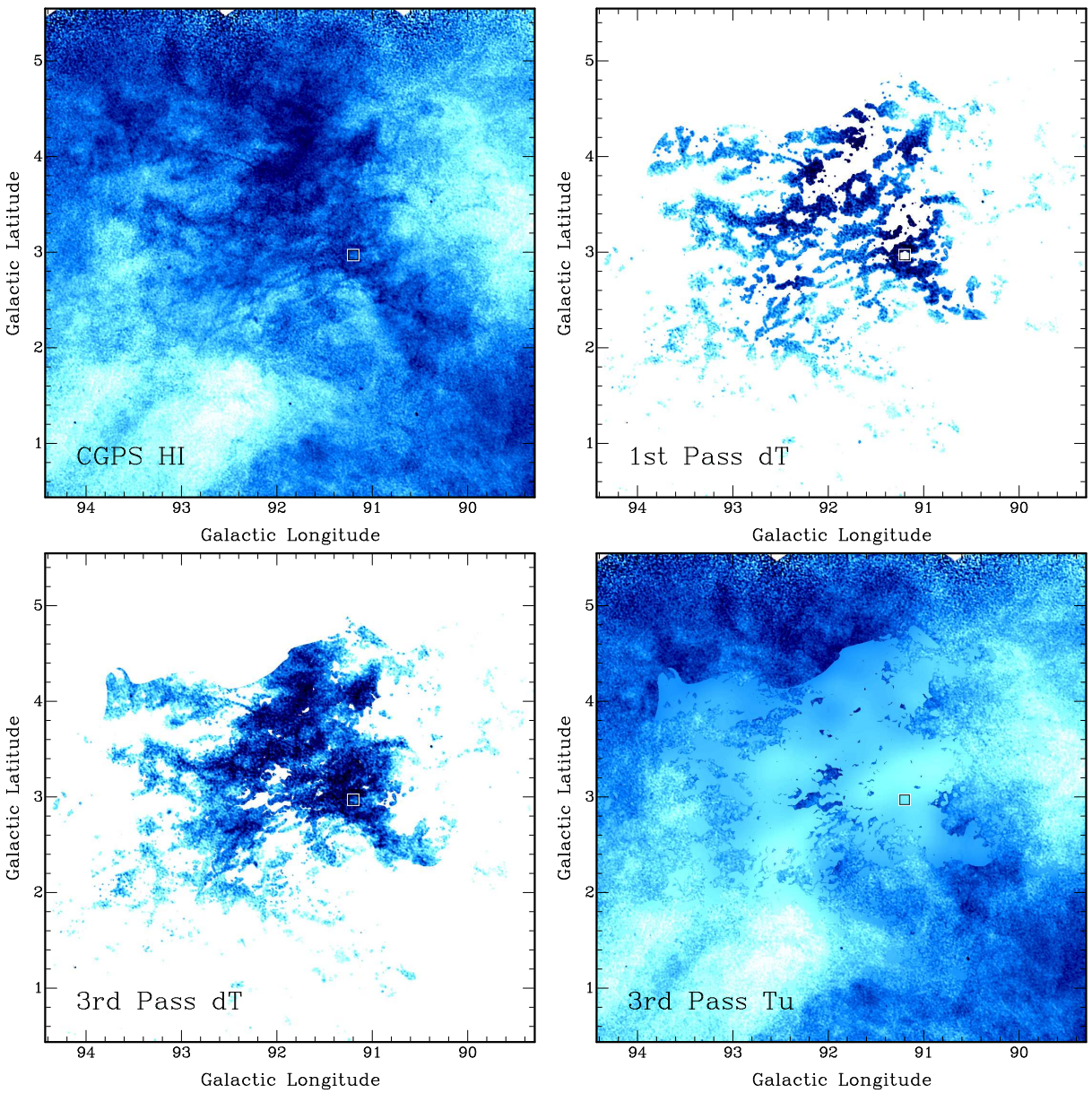

Fig. 7.— (ℓ, b) channel maps illustrating multiple-pass extraction of a large HISA complex in the CGPS MK2 mosaic cube at at -3 km s^{-1}. Shown are H I brightness, first-pass ΔT, third-pass ΔT, and corresponding T_U. Intensity ranges are 0 to +130 K for the first and last panels and -65 to 0 K for the two ΔT maps. The small $6' \times 6'$ box ($\ell = 91.20°, b = +2.97°$) marks the area from which the spectra in Figure 8 were extracted. The feature extraction is truncated for $b \gtrsim +4.5°$ due to $T_U < 70$ K (see §2.4).

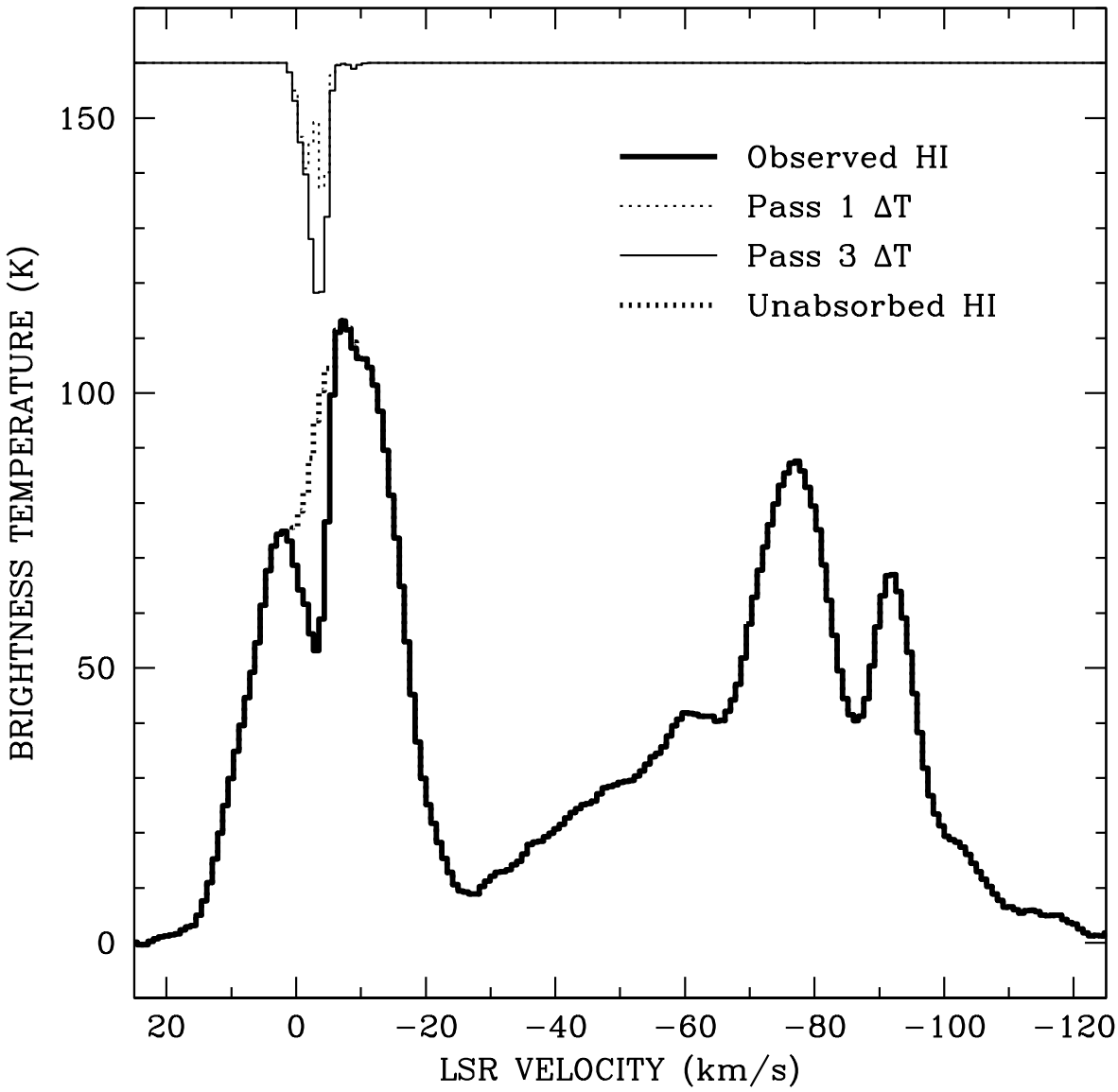

Fig. 8.— Spatially-averaged velocity spectra illustrating multiple-pass HISA extraction. The spectra are extracted from the $6' \times 6'$ box marked in Figure 7 at $(\ell = 91.20°, b = +2.97°)$. Shown are H I emission, first-pass HISA ΔT, third-pass HISA ΔT, and third-pass HISA T_U. For clarity, the ΔT zero-points have been shifted to 160 K.

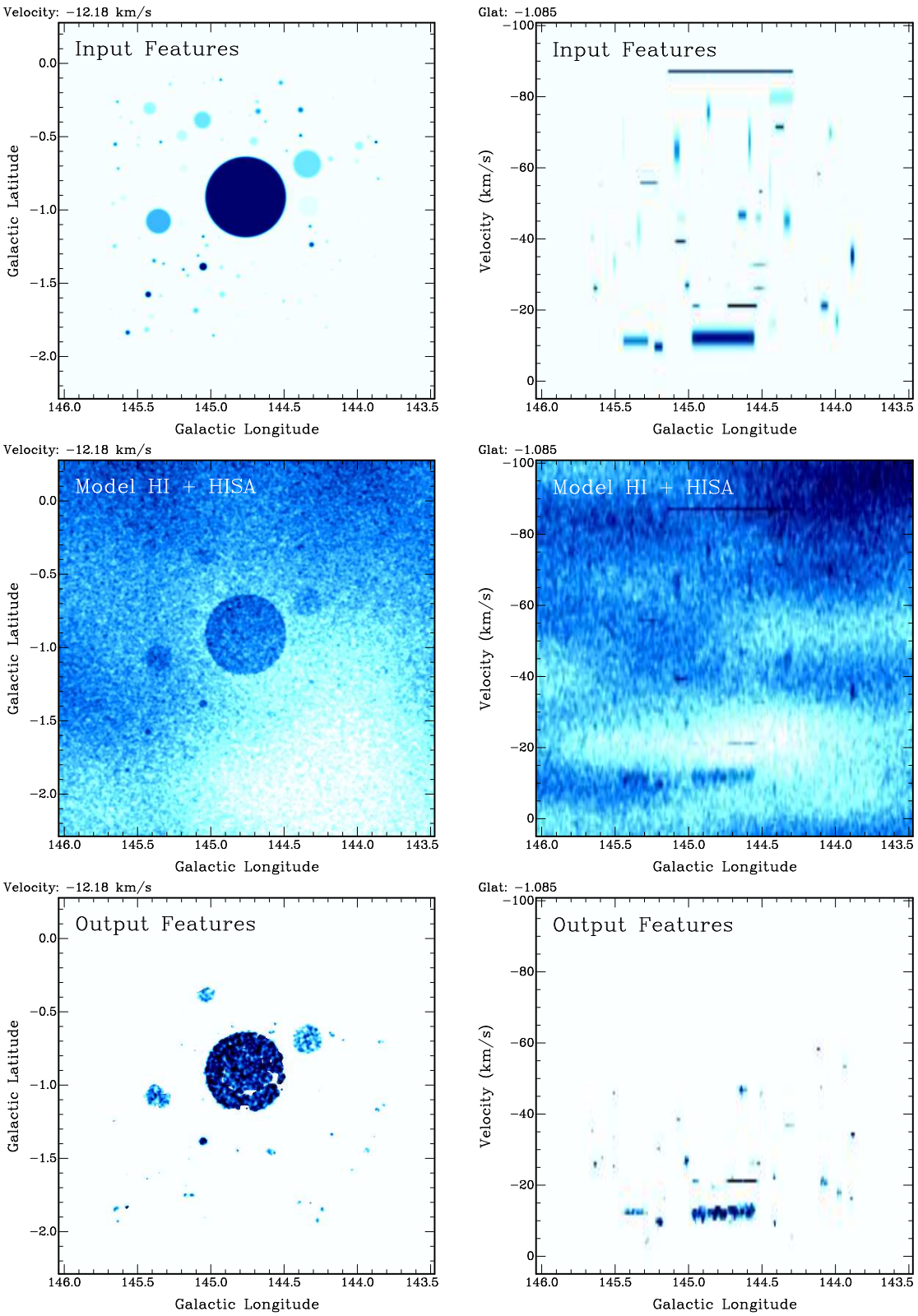

Fig. 9.— (ℓ, b) and (ℓ, v) slices of sample model H I data, showing the input HISA "hockey puck" amplitude ΔT_{in}, the noisy background emission field $T_{U\,in}$ with pucks added, and the extracted HISA amplitude ΔT_{out} after the 3rd identification pass. Intensity ranges are −40 to 0 K for the ΔT maps and 0 to 120 K for the H I maps, from black to white.

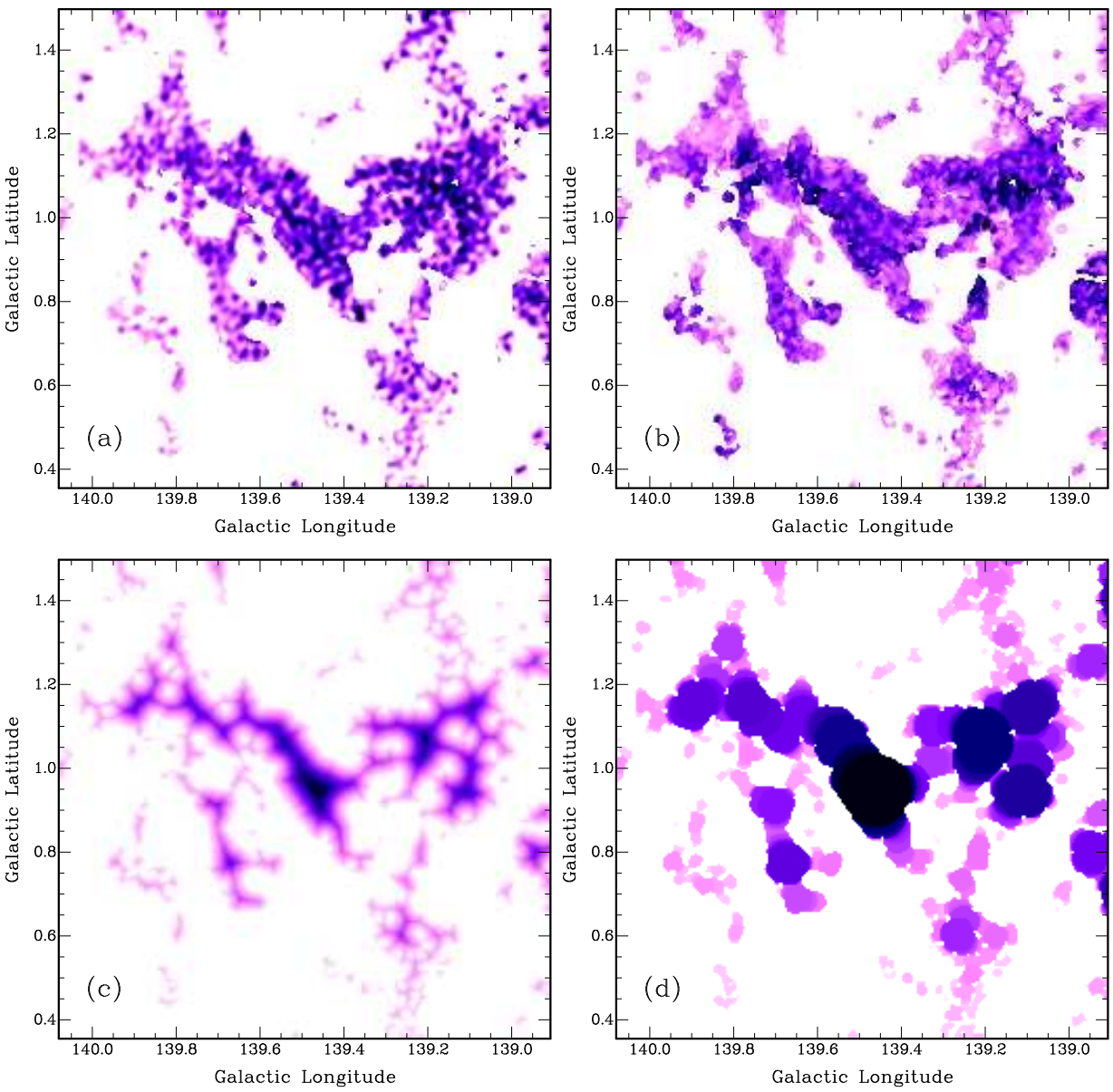

Fig. 10.— Channel maps illustrating velocity width and angular width measures: *(a)* HISA absolute amplitude $|\Delta T|$; *(b)* Δv, the line full width at half maximum; *(c)* $2 \times \theta_{\rm off}$, where $\theta_{\rm off}$ is the offset to the nearest HISA feature edge; and *(d)* $\Delta\theta$, the angular width obtained from $2\,\theta_{\rm off}$ values imposed out to a radius $\theta_{\rm off}$. Intensity ranges are linear, from white to black, for 0 K $\leq |\Delta T| \leq$ 40 K, 0 km s^{-1} $\leq \Delta v \leq$ 6 km s^{-1}, $0' \leq 2\theta_{\rm off} \leq 10'$, and $0' \leq \Delta\theta \leq 10'$. See §3.3.2 for further details.

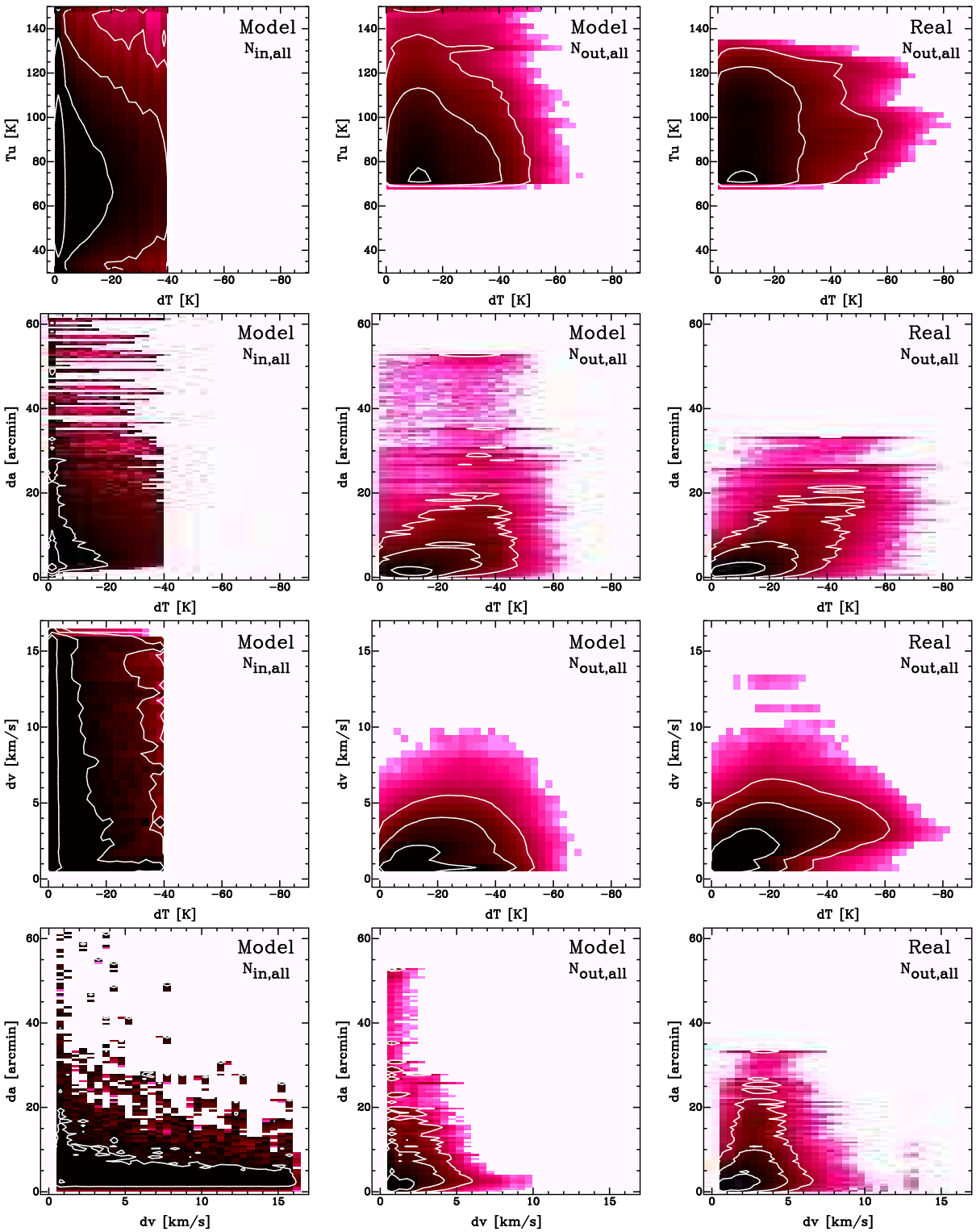

Fig. 11.— 2-D projections of 4-D property histograms of input voxels (model $N_{in,all}$), extracted voxels (model $N_{out,all}$), and observed CGPS HISA (real $N_{out,all}$). Axis labels are "dT" = ΔT, "Tu" = T_U, "da" = $\Delta\theta$, and "dv" = Δv. Counts were summed along the orthogonal axes, so the full distributions are visible. No significant trends in $(T_U, \Delta\theta)$ or $(T_U, \Delta v)$ were found. The intensity scale is logarithmic from 1 count (light) to 1 million counts (dark). Contours mark counts of 10^3, 10^4, 10^5, 10^6, and 10^7 in all panels except the $(\Delta T, \Delta\theta)$ and $(\Delta v, \Delta\theta)$ maps of the model $N_{in,all}$, where 10^3 and 10^4 are omitted.

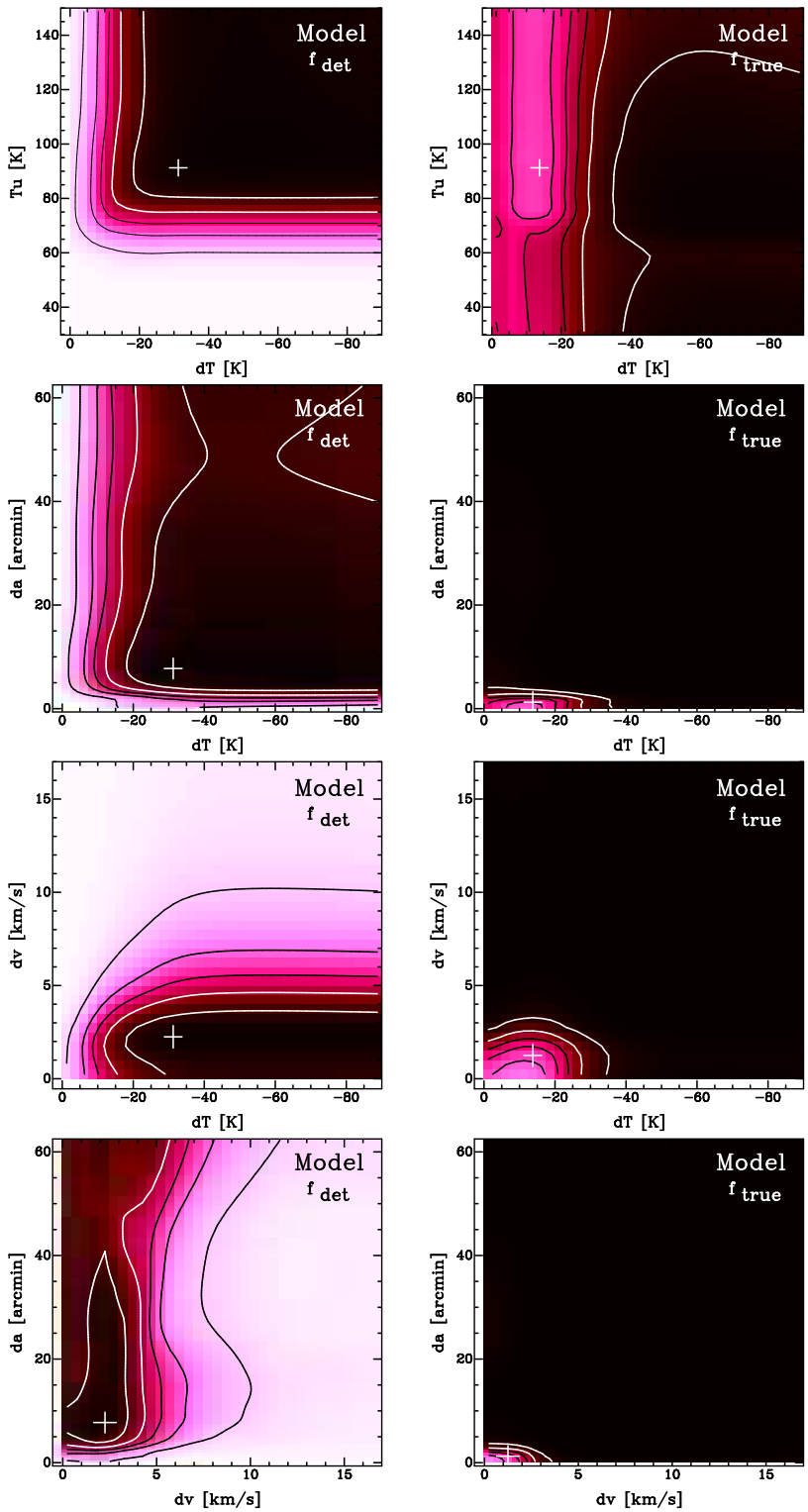

Fig. 12.— 2-D slices through the 4-D throughput f_{det} and true fraction f_{true} histograms. The f_{det} slices intersect at a common position marked with a cross. The f_{true} slices intersect at a different common position, also marked. The intensity scale is linear, from 0.0 (white) to 1.0 (black). Black contours mark values of 0.1, 0.3, and 0.5; white contours mark values of 0.7 and 0.9. Axis labels are as in Figure 11.

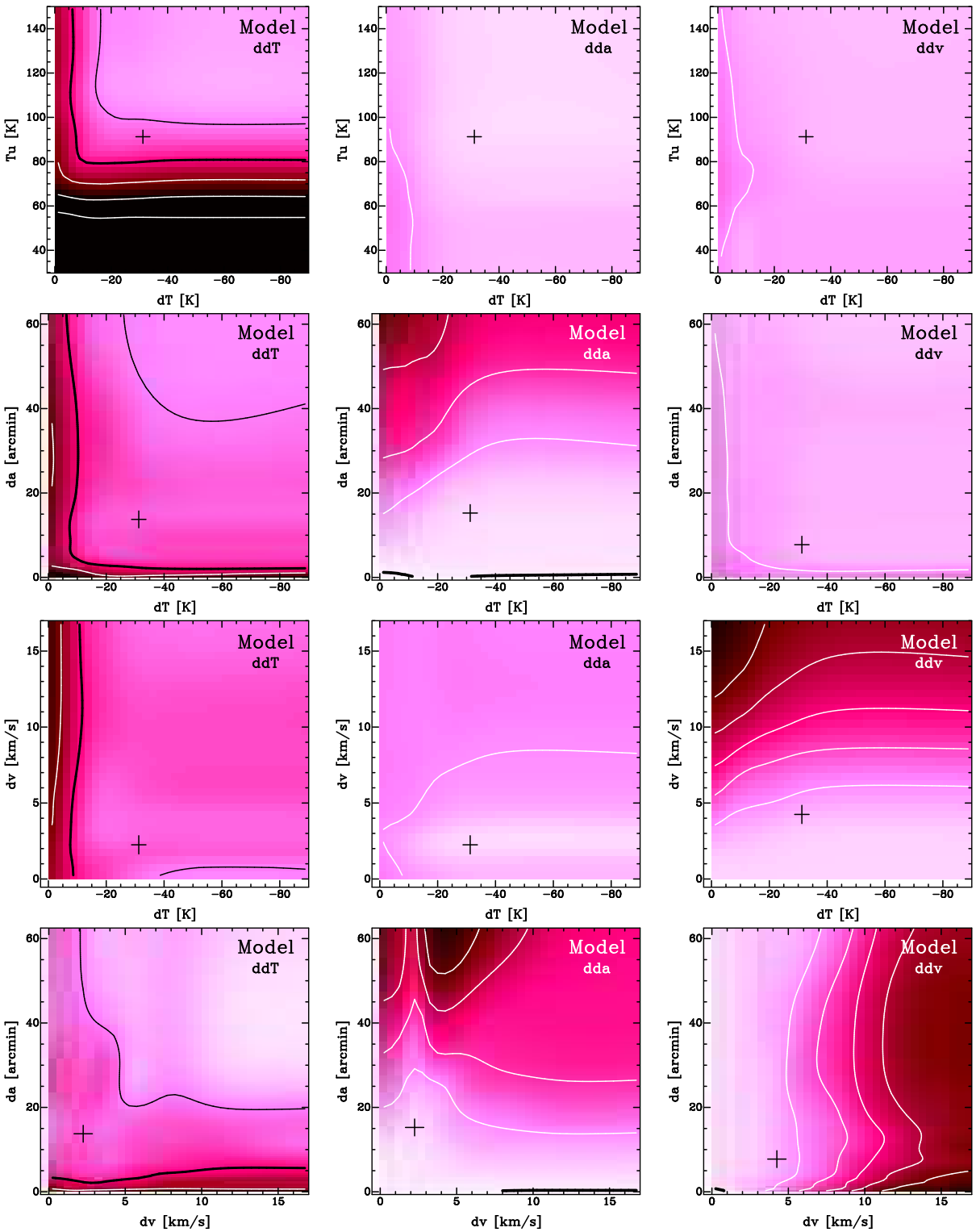

Fig. 13.— 2-D slices through the 4-D drift histograms $\Delta\Delta T$, $\Delta\Delta\theta$, and $\Delta\Delta v$. As in Figure 12, crosses mark slice intersections for each 4-D drift measure. The intensity scale is linear, from negative (black) to positive (white). The intensity ranges are $-10\ \mathrm{K} < \Delta\Delta T < +10\ \mathrm{K}$, $-50' < \Delta\Delta\theta < +1'$, and $-13\,\mathrm{km\,s^{-1}} < \Delta\Delta v < +1\,\mathrm{km\,s^{-1}}$. Where present, a thick black contour marks zero drift. Thinner contours mark positive (black) and negative (white) drifts at intervals of 5 K, $10'$, and $2\,\mathrm{km\,s^{-1}}$, respectively. Axis labels are as in Figure 11.

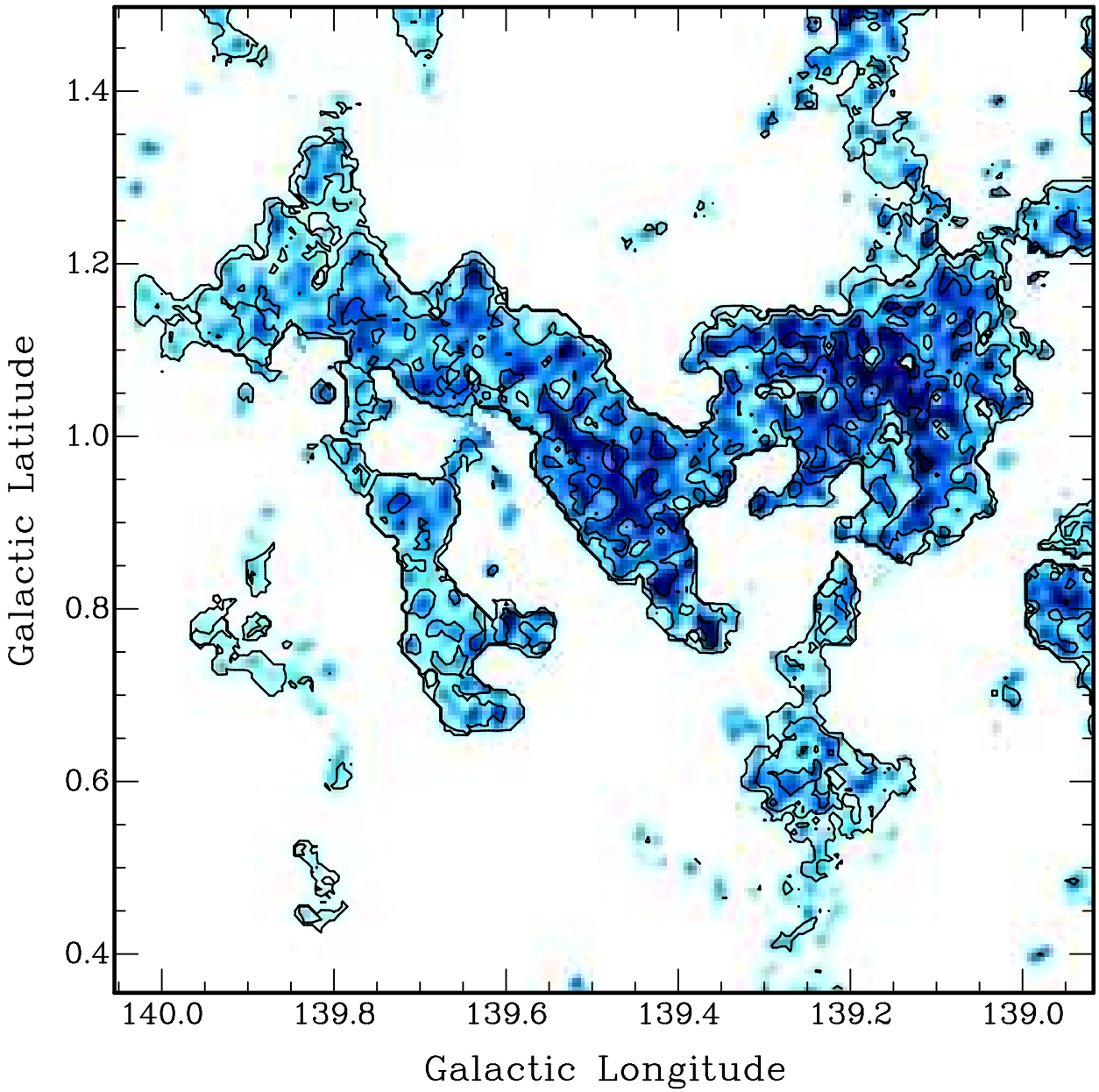

Fig. 14.— Sample extracted HISA $\Delta T(\ell, b)$ map for a single velocity, showing contours of f_{true} = 0.682690, 0.954500, 0.997300, and 0.999937, which correspond to reliability thresholds of 1, 2, 3, and 4 σ if Gaussian statistics apply. The maximum f_{true} in this map is 0.999999, which is equivalent to 4.90 σ. The region shown is the same as in Figure 10.