CDF/PUB/EXOTIC/PUBLIC/10574
July 21, 2011

Search for the Higgs Boson Produced in Association with Top Quarks Using $7.5\,\mathrm{fb}^{-1}$

Jon Wilson Homer Wolfe Jake Connors Richard Hughes Brian Winer
The Ohio State University

Abstract

We present a new Higgs boson search analysis using the $t\bar{t}H$ associated production channel in the lepton plus jets final state. This analysis has a similar final state to the existing $WH \to \ell\nu b\bar{b}$ analysis, so we have used much of the existing WH machinery. We exploit the high jet and high b-jet multiplicity in these events to both select a sample expected to contain $t\bar{t}H$ and to maintain orthogonality with other Higgs boson searches using the lepton plus jets final state. We search for a Higgs boson in the range $100\,\mathrm{GeV}/c^2 < m_H < 170\,\mathrm{GeV}/c^2$, using neural networks optimized for each mass point independently. Using $7.5\,\mathrm{fb}^{-1}$ of data, we obtain an expected (observed) limit on the Higgs boson production cross section of 11.7 (22.9) times the expected Standard Model value for a Higgs boson mass of $115\,\mathrm{GeV}/c^2$.

Preliminary Results

Contents

1 Introduction

This note details a low mass Higgs boson search analysis using the process $t\bar{t}H$. The target sample is one lepton plus missing transverse energy plus at least 4 jets, with at least two of the jets b tagged. Although significant acceptance comes from the Higgs boson decay into two b quarks, there is no explicit requirement for this decay in this search. The overwhelming background to the process is standard model $t\bar{t}$ production. Figure 1 shows a Feynmann diagram of the $t\bar{t}H$ process, assuming the Higgs decays to two b quarks.

The final state for this process is lepton plus jets, and as a result we use many of the techniques which have been developed for both the top group as well as the WH group. The basic strategy is to require a well identified high p_T lepton, significant missing transverse energy, and at least 4 jets. We also use two different algorithms to identify jets originating from a b quark. The SECVTX algorithm[2] identifies displaced vertices, and the Jet Probability algorithm[3] uses track impact parameters. We define 5 tagging samples, composed of various combinations and numbers of jets tagged by these two algorithms. For the purposes of this note, we have combined all of the 2-tag categories together and all of the 3-tag categories together for validation plots, but the categories are kept separate in the analysis.

Once our samples have been defined, we pass the selected events through a novel ensemble discriminant. This discriminant has a number of useful and interesting fea-

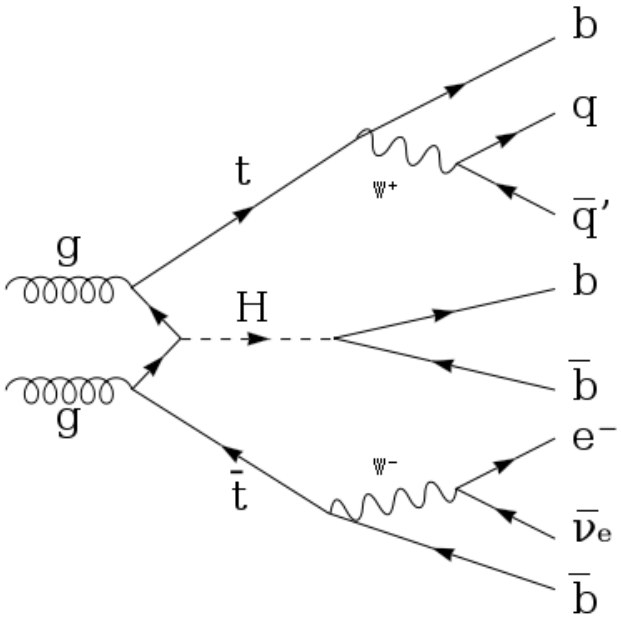

Figure 1: The $t\bar{t}H$ process.

tures which we describe in detail below. We use this discriminant to set 95 % C.L. upper limits on the Higgs boson production cross section.

2 Monte Carlo Samples

Our Higgs boson signal model comes from the Monte Carlo samples generated with PYTHIA[5]. These Higgs boson samples were generated for a range of Higgs boson masses from $100\,\mathrm{GeV}/c^2$ to $150\,\mathrm{GeV}/c^2$ in incremements of $5\,\mathrm{GeV}/c^2$ as well as one sample at $170\,\mathrm{GeV}/c^2$. We include the $170\,\mathrm{GeV}/c^2$ sample to explore the sensitivity of this analysis at high Higgs boson masses. The W and Z plus light-flavor and heavy-flavor jet processes are modeled using ALPGEN version 2.10[4] showered through PYTHIA. Likewise, the single-top contribution is modeled using parton-level events generated by MadEvent[6] and showered through PYTHIA. The rest of the background processes, including the $t\bar{t}$, WW, WZ, and ZZ processes were generated with PYTHIA. For backgrounds involving a top quark, the top mass was set to $172.5\,\mathrm{GeV}/c^2$.

3 Data Samples

We use data taken by the CDF detector between February 2002 and March 2011, corresponding to an integrated luminosity of $7.5\,\mathrm{fb}^{-1}$. We use data taken with three different triggers: the high p_T electron trigger, the plug electron trigger, and the high p_T muon trigger. We use a standard CDF luminosity calculation[8], including corrections for the trigger system.

4 Event Selection

The basic strategy is the same as in the WH search[1], with the exception of requiring a higher jet multiplicity. We require a well identified high p_T lepton, significant missing energy, and at least 4 jets.

4.1 Lepton Identification

We use standard CDF definitions[1] for our lepton types: high p_T central electrons, high p_T plug electrons, and high p_T muons.

4.2 Missing Transverse Energy

$\not\!\!E_T$ is calculated according to standard CDF calculations[1] including corrections for vertex position, for the presence of muons, and for corrections to jet energies. We then select events with corrected $\not\!\!E_T$ above $20\,\mathrm{GeV}$ (above $25\,\mathrm{GeV}$ for plug electrons).

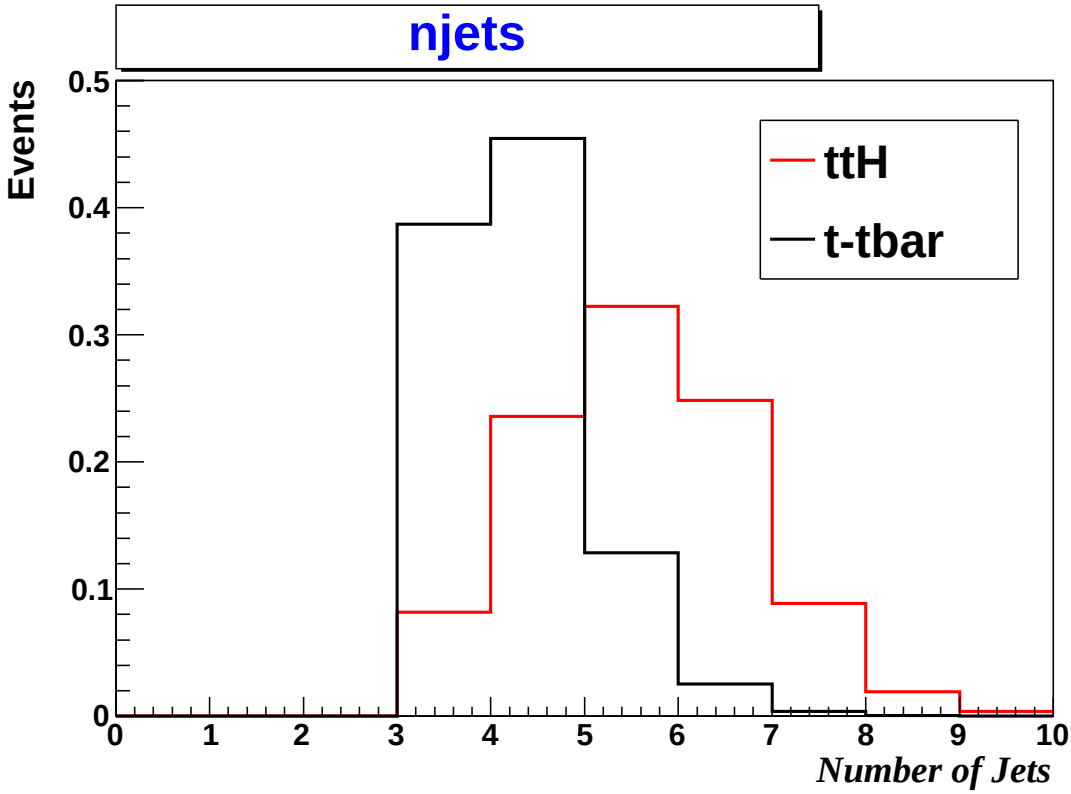

Figure 2: The distribution of the number of jets after all other cuts for $t\bar{t}$ events versus $t\bar{t}H$ events.

4.3 Jet Selection

All jet energies are corrected according to standard CDF prescriptions. We then use jets with $E_t \geq 20\,\text{GeV}$ and $|\eta| \leq 2.0$[1].

Figure 2 shows the distribution of the number of jets for $t\bar{t}$ events versus $t\bar{t}H$ events. For our final signal sample selection, we require that events have at least 4 jets. We further divide this sample into two categories, requiring exactly 4 jets and at least 5 jets.

4.4 b Tagging

Our signal sample will contain 2 or 4 bottom quarks, depending on the Higgs boson decay. In addition, we can expect some tagging acceptance from τ or charm decays of the W (either from the decay of the top quarks or $H \to WW^{\star}$). This leads us to require multiple tags in the final state. In addition, since different numbers of tags will in general have different signal to background ratios, we divide our sample based on the the number and types of tags observed. Events appear only once in each category,

and are placed into the highest signal to background ratio category they satisfy. The following list is arranged in order of signal to background ratio:

STSTST These are events with at least 3 separate jets that are tagged by the SECVTX algorithm.

STSTJP These are events with exactly 2 jets that are tagged by the SECVTX algorithm, and at least 1 additional jet that is tagged by the Jet Probability algorithm.

STJPJP These are events with exactly 1 jet that is tagged by the SECVTX algorithm, and at least 2 additional jets that are tagged by the Jet Probability algorithm.

STST These are events with exactly 2 separate jets that are tagged by the SECVTX algorithm.

STJP These are events with exactly 1 jet that is tagged by the SECVTX, and exactly 1 additional jet that is tagged by the Jet Probability algorithm.

We will use "multiple-tagged" to refer to the combination of all of the above 5 separate tagging categories. We will used "2-tag" to refer to the combination of the STST and STJP categories and "3-tag" to refer to the combination of the STSTST, STSTJP, and STJPJP categories.

4.5 Predicted Backgrounds

The overwhelming background in this analysis is $t\bar{t}$ events, predicted to be more than 85 % of the selected sample. However, we will consider all of the backgrounds in the search, following the same methodology for background estimation as the WH search at CDF.

The backgrounds in order of size (summed over all 5 b-tagging categories in the 5-jet bin):

$\boldsymbol{t\bar{t}}$ **+ jets** is modeled using PYTHIA Monte Carlo. This sample is expected to comprise ~ 90 % of the ≥ 5 jets multiple-tagged sample.

$\boldsymbol{Wb\bar{b}}$ is modeled with ALPGEN v2 + PYTHIA Monte Carlo.

Non-$\boldsymbol{W}$ is estimated according to the standard CDF prescriptions, by reversing any two of the lepton identification cuts.

$\boldsymbol{W}$ **+ Charm** includes both $Wc\bar{c}$ and Wc, and is modeled with ALPGEN v2 + PYTHIA Monte Carlo.

Mistags includes both W and Z plus light flavor jets. This contribution is estimated according to the standard CDF prescriptions, by applying the standard estimation of the tagging rate on light flavor jets to the W/Z + light flavor Monte Carlo.

Sample	$\boldsymbol{N}_{\text{jets}} == \mathbf{4}$	$\boldsymbol{N}_{\text{jets}} \geq \mathbf{5}$
DiTop	80.02 ± 10.05	39.35 ± 4.90
STopT	0.38 ± 0.05	0.13 ± 0.02
STopS	0.55 ± 0.06	0.19 ± 0.02
Wbb	3.76 ± 0.99	1.72 ± 0.45
Wcc	0.74 ± 0.23	0.40 ± 0.12
Wcj	0.36 ± 0.12	0.16 ± 0.05
Zjets	0.14 ± 0.01	0.07 ± 0.01
WW	0.17 ± 0.02	0.05 ± 0.01
WZ	0.08 ± 0.01	0.03 ± 0.00
ZZ	0.00 ± 0.00	0.00 ± 0.00
Non-W	0.89 ± 2.83	0.91 ± 2.79
Mistags	0.36 ± 0.13	0.18 ± 0.07
Total Prediction	87.45 ± 10.49	43.19 ± 5.66
ttH120	0.14 ± 0.01	0.68 ± 0.04
Observed	73	48

Table 1: Background from various sources compared to observed data, for the 3-tag categories.

Single top includes both s- and t-channel contributions and is modeled using MadEvent + PYTHIA Monte Carlo.

Diboson includes WW, WZ, and ZZ, and is modeled using PYTHIA Monte Carlo.

$\boldsymbol{Z}$ **+ jets** includes both $Zb\bar{b}$ and $Zc\bar{c}$ and is modeled with ALPGEN v2 + PYTHIA Monte Carlo.

5 Signal Discrimination

In order to discriminate the signal from the backgrounds, we employ an ensemble of neural networks. For each Higgs boson mass candidate, we train an ensemble of 1000 neural networks to classify $t\bar{t}$ and $t\bar{t}H$. We then combine the output of the 1000 constituent neural networks using a method called "Supra-Bayesian". The constituent neural networks were trained through 70 epochs, with 10 input variables, 15 hidden nodes in a single hidden layer, and one output node. The 10 input variables were chosen at random for each constituent neural network from the list of 21 candidate input variables below.

Sample	$N_{\text{jets}} == 4$	$N_{\text{jets}} \geq 5$
DiTop	493.82 ± 40.47	168.23 ± 13.19
STopT	4.83 ± 0.38	0.95 ± 0.07
STopS	4.35 ± 0.29	0.90 ± 0.06
Wbb	33.80 ± 10.29	9.35 ± 2.96
Wcc	9.92 ± 3.47	3.37 ± 1.22
Wcj	4.94 ± 1.72	1.37 ± 0.49
Zjets	2.25 ± 0.22	0.70 ± 0.07
WW	1.64 ± 0.27	0.55 ± 0.09
WZ	0.71 ± 0.07	0.21 ± 0.02
ZZ	0.10 ± 0.01	0.02 ± 0.00
Non-W	17.87 ± 11.57	5.87 ± 4.64
Mistags	8.49 ± 2.42	2.71 ± 0.96
Total Prediction	582.72 ± 43.58	194.23 ± 14.38
ttH120	0.27 ± 0.01	0.74 ± 0.04
Observed	561	210

Table 2: Background from various sources compared to observed data, for the 2-tag tagging categories.

5.1 Ensemble Method

A subsample of the $t\bar{t}$ and $t\bar{t}H$ Monte Carlo samples is identified as a "testing" sample, and the constituent neural networks are evaluated for all events in the testing samples, producing a background and a signal output shape for each neural network. These shapes are stored as histograms along with the neural networks. To evaluate the ensemble on a novel event, we evaluate each constituent neural network on the event, and look up the fraction of expected background (B) and signal (S) events that would have an output in the same bin of the stored histograms. The output of the ensemble is then the simple average, over all 1000 constituent neural networks, of $\frac{S}{S+B}$.

This technique provides a discriminant that is only marginally more powerful than a single neural network. Nonetheless, because some of our input variables (listed below) have discrete values, a single neural network output would be very choppy, with multiple very sharp peaks. While this has a minimal impact on the *expected* sensitivity of the analysis, the *observed* limit fluctuates quite widely (much more than the uncertainty bands on the expected limit would suggest) upon retraining when using a single neural network. The ensemble averaging process smooths the output shape of the discriminant and brings this observed limit fluctuation under control.

5.2 Discriminating Variables

Since the overwhelming background to the $t\bar{t}H$ process is $t\bar{t}$, we look for variables which can distinguish these two processes. The variables we consider are:

- $\not\!\!E_T$ corrected according to standard procedures
- Lepton p_T
- Lepton η
- Maximum Jet E_T
- Mean Jet E_T
- Number of jets
- Event sum Mass
- Event sum E_T
- Minimum ΔR between tagged jets
- Jet 2 E_T
- Jet 3 E_T
- ΔR between lepton and the nearest jet
- ΔR between lepton and the nearest tagged jet
- ΔR between lepton and the $\not\!\!E_T$
- Maximum Tagged Jet E_T
- W Transverse Mass
- Lepton plus nearest jet mass
- Jet 1 η
- Summed E_T of tight jets
- Minimum dijet mass
- Dijet mass of untagged jets

Validation plots of two of these variables are shown in figures 3 and 4. The final discriminants for $m_H = 115\,\mathrm{GeV}/c^2$ are shown in figure 5. Validation plots for the other input variables and for the other outputs are shown on the public web page for the analysis.

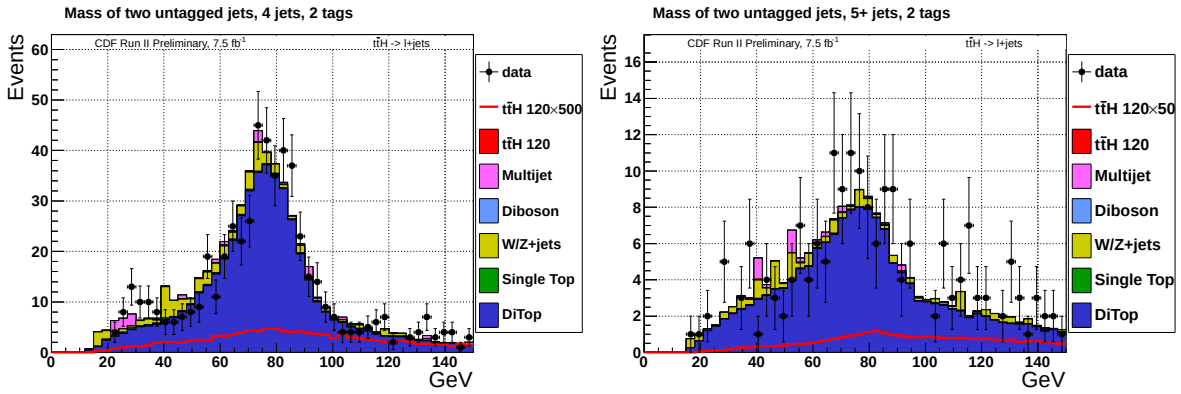

Figure 3: Mass of leading untagged jets showing W boson peak, in 4 and 5+ jets, with 2 tags.

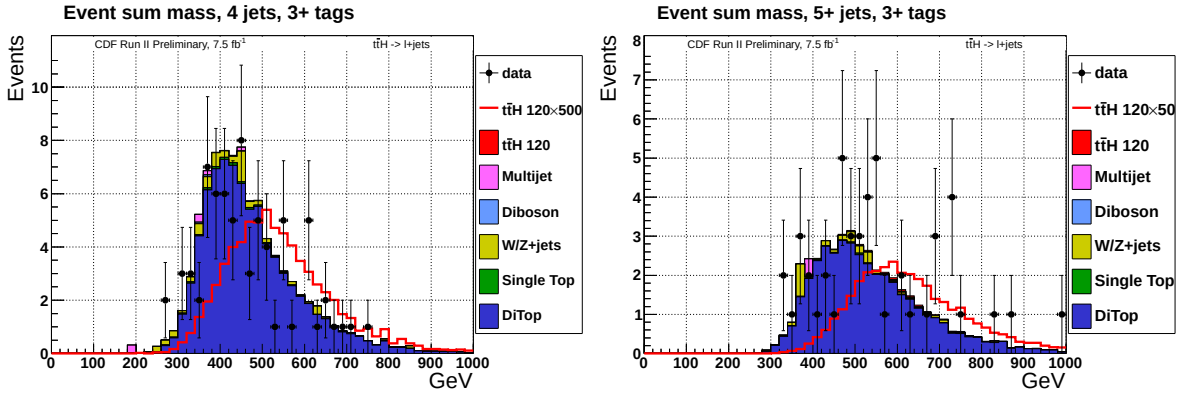

Figure 4: Mass of vector sum of all objects in the event, in 4 and 5+ jets, with 3+ tags.

6 Systematic Uncertainties

This analysis includes many of the same systematics that are used in the WH analysis. The major systematics include the uncertainties on the process cross sections and the jet energy scale (JES) systematic, which can strongly effect the number of jets in an event. This JES systematic not only affects the rate of the various processes, but also the shape of the discriminants.

Other important systematics are the uncertainty on the b-tag scale factors which account for the difference between the b-tag rates for Monte Carlo and for data, the uncertainty on the tagging rate for light flavor jets, the uncertainty on the measurement of the luminosity delivered to CDF, and the uncertainty on the amount of initial and final state radiation (ISR/FSR), which we apply to both the dominant $t\bar{t}$ background and to the signal.

Tables 3 and 4 summarize the various systematics applied to the $t\bar{t}$ background as well as the signal. Because the total rate of the other backgrounds is so small, we do

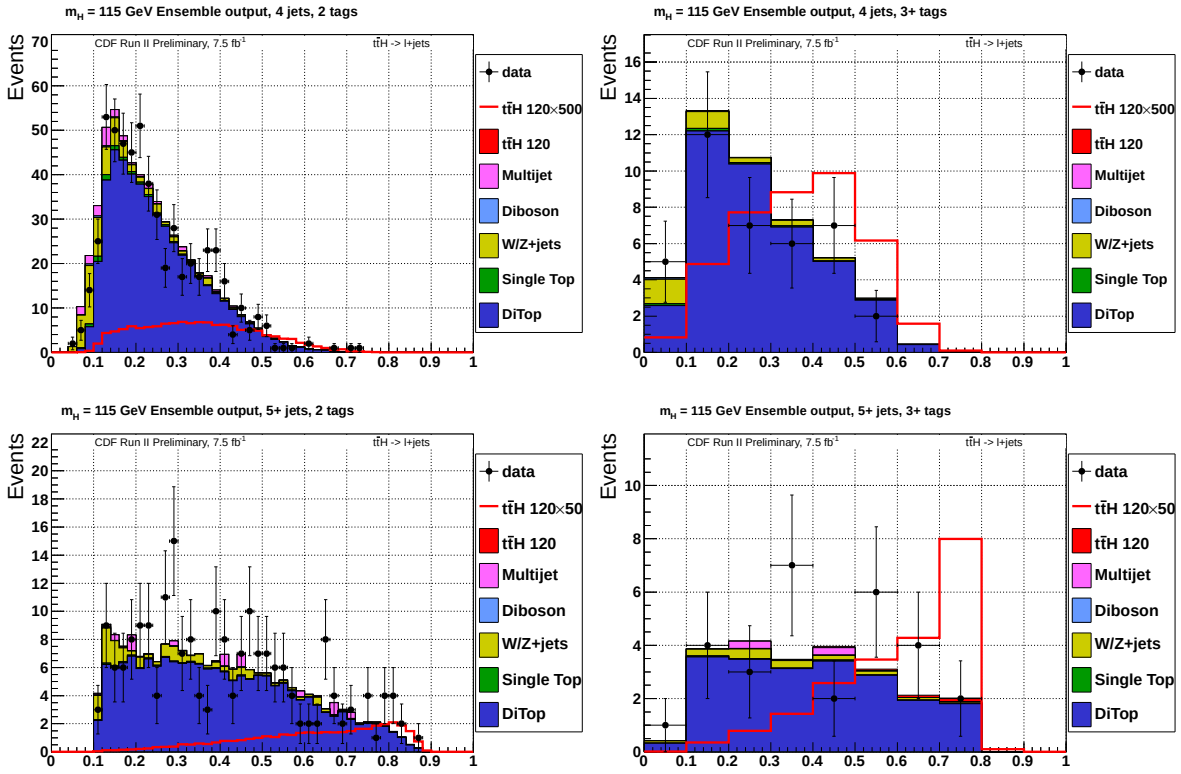

Figure 5: Discriminant outputs, discriminants trained at $m_H = 115\,\mathrm{GeV}/c^2$

4 jets	STJP		STJPJP		STST		STSTJP		STSTST	
	$t\bar{t}$	$t\bar{t}H$	$t\bar{t}$	$t\bar{t}H$	$t\bar{t}$	$t\bar{t}H$	$t\bar{t}$	$t\bar{t}H$	$t\bar{t}$	$t\bar{t}H$
$t\bar{t}H$ cross section	0	10	0	10	0	10	0	10	0	10
$t\bar{t}$ cross section	10	0	10	0	10	0	10	0	10	0
Tevatron luminosity	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8
CDF luminosity	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4
b-tag scale factor	$^{+1.4}_{-2.5}$	$^{-2.9}_{-2.0}$	$^{+3.3}_{-1.5}$	$^{+0.3}_{+0.3}$	$^{+7.3}_{-9.4}$	$^{+6.7}_{-2.0}$	$^{+8.3}_{-8.8}$	$^{+7.0}_{-7.7}$	$^{+11}_{-12}$	$^{+11}_{-16}$
light jet tag rate	$^{+1.7}_{-2.0}$	$^{-0.4}_{-1.5}$	$^{+10}_{-11}$	$^{-1.1}_{-5.7}$	$^{-1.2}_{+2.7}$	$^{+2.7}_{+3.7}$	$^{+7.6}_{-7.4}$	$^{+1.7}_{+2.4}$	$^{+3.3}_{-5.1}$	$^{+1.6}_{+0.2}$
jet energy scale	$^{+3.8}_{-5.1}$	$^{-13}_{+6.7}$	$^{+2.5}_{-4.5}$	$^{0.0}_{0.0}$	$^{+4.2}_{-4.8}$	$^{-5.9}_{+5.9}$	$^{+2.5}_{-3.8}$	$^{-12}_{0.0}$	$^{+3.3}_{-4.4}$	$^{-12}_{0.0}$
initial- and final-state radiation	$^{-1.8}_{-1.0}$	$^{-0.1}_{+0.1}$	$^{-1.3}_{+2.3}$	$^{-0.5}_{+0.5}$	$^{-3.8}_{-1.3}$	$^{+0.2}_{-0.2}$	$^{-4.4}_{-1.1}$	$^{+0.0}_{-0.0}$	$^{-2.9}_{-3.5}$	$^{-0.2}_{+0.2}$

Table 3: Systematic uncertainties in 4 jets. The b-tag scale factor, light jet tag rate, jet energy scale, and initial- and final-state radiation systematics are all shape+rate systematics, but only the rate portion is shown here.

not show the effects of the systematics that we apply to them. Uncertainties shown are relative, in percent, and are symmetric unless otherwise indicated.

5 jets	STJP		STJPJP		STST		STSTJP		STSTST	
	$t\bar{t}$	$t\bar{t}H$	$t\bar{t}$	$t\bar{t}H$	$t\bar{t}$	$t\bar{t}H$	$t\bar{t}$	$t\bar{t}H$	$t\bar{t}$	$t\bar{t}H$
$t\bar{t}H$ cross section	0	10	0	10	0	10	0	10	0	10
$t\bar{t}$ cross section	10	0	10	0	10	0	10	0	10	0
Tevatron luminosity	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8
CDF luminosity	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4
b tag scale factor	$^{+1.8}_{-3.5}$	$^{-0.4}_{+2.7}$	$^{+4.5}_{-4.1}$	$^{-1.3}_{-1.6}$	$^{+8.2}_{-6.8}$	$^{+2.5}_{-5.0}$	$^{+9.7}_{-7.7}$	$^{+5.9}_{-5.5}$	$^{+11}_{-16}$	$^{+9.9}_{-13}$
light jet tag rate	$^{+1.3}_{-2.9}$	$^{-7.5}_{+1.8}$	$^{+18}_{-8.9}$	$^{+4.3}_{-6.6}$	$^{-0.2}_{+2.6}$	$^{-2.0}_{+1.0}$	$^{+8.2}_{-8.7}$	$^{+2.5}_{-2.2}$	$^{+8.1}_{-3.4}$	$^{+1.3}_{-0.5}$
jet energy scale	$^{+19}_{-16}$	$^{+7.5}_{-7.5}$	$^{+17}_{-15}$	$^{+7.1}_{-14}$	$^{+18}_{-17}$	$^{+7.0}_{-4.7}$	$^{+16}_{-16}$	$^{+6.7}_{-3.3}$	$^{+15}_{-15}$	$^{-2.7}_{-8.1}$
initial- and final-state radiation	$^{+10}_{-1.2}$	$^{-0.0}_{+0.0}$	$^{+14}_{-1.0}$	$^{-0.2}_{+0.2}$	$^{+8.2}_{-6.5}$	$^{+0.0}_{-0.0}$	$^{+12}_{-5.1}$	$^{-2.1}_{+2.1}$	$^{+14}_{-2.0}$	$^{-1.9}_{+1.9}$

Table 4: Systematic uncertainties in 5 jets. The b-tag scale factor, light jet tag rate, jet energy scale, and initial- and final-state radiation systematics are all shape+rate systematics, but only the rate portion is shown here.

7 Results

Using the outputs of the final event discriminants described above, we observe no evidence of a $t\bar{t}H$ signal and proceed to set limits on the Higgs boson production cross section for this channel. We use the MCLimit machinery[7] to produce median, $\pm 1\sigma$, and $\pm 2\sigma$ expected limits, along with the observed limits. MCLimit uses a Bayesian technique involving many pseudoexperiments to marginalize the systematic uncertainties and find the expected and observed lower bounds on the Higgs boson production cross section. This is done for $100\,\mathrm{GeV}/c^2 \leq m_H \leq 150\,\mathrm{GeV}/c^2$ in steps of $5\,\mathrm{GeV}/c^2$, as well as for $170\,\mathrm{GeV}/c^2$. We use 10 different MCLimit channels: one for each tagging category, separated into 4 jets and ≥ 5 jets.

7.1 Observed and Expected Limits

The expected and observed limits are shown in table 6 and figure 6. The limits for the 4 jet bin alone and for the ≥ 5 jet bin alone are shown in figure 7.

Higgs Boson Mass	$t\bar{t}H$ **Cross section (fb)**
100	7.99
110	6.28
120	4.94
130	3.88
140	3.05

Table 5: Cross sections at $\sqrt{s} = 1.96\,\text{TeV}$ for $t\bar{t}H$

m_H	Obs	-2σ	-1σ	Exp	$+1\sigma$	$+2\sigma$
100	**16.3**	4.5	6.2	**8.9**	13.0	18.3
105	**19.0**	4.8	6.7	**10.0**	14.5	19.7
110	**18.0**	5.4	7.2	**10.3**	14.8	21.3
115	**22.9**	6.0	8.3	**11.7**	16.9	24.1
120	**27.4**	6.3	8.7	**12.7**	19.1	26.7
125	**25.6**	7.2	9.7	**14.0**	20.4	30.1
130	**26.6**	8.5	11.4	**16.6**	24.0	33.1
135	**34.9**	9.7	13.6	**18.5**	27.3	39.1
140	**33.1**	10.6	13.9	**19.7**	29.0	42.5
145	**40.6**	11.5	15.6	**21.5**	30.7	44.5
150	**47.2**	11.9	16.6	**22.4**	33.2	46.7
170	**56.6**	17.8	23.1	**32.3**	46.6	62.4

Table 6: Observed and expected limits, for all tagging categories and both jet bins combined.

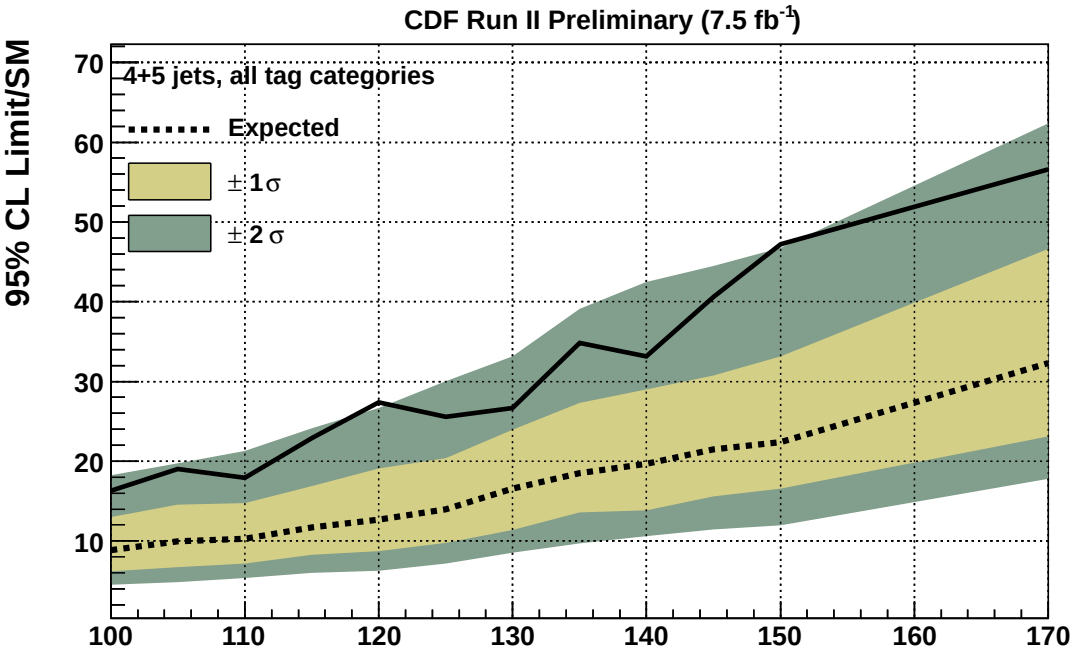

Figure 6: Expected and observed limits for this analysis.

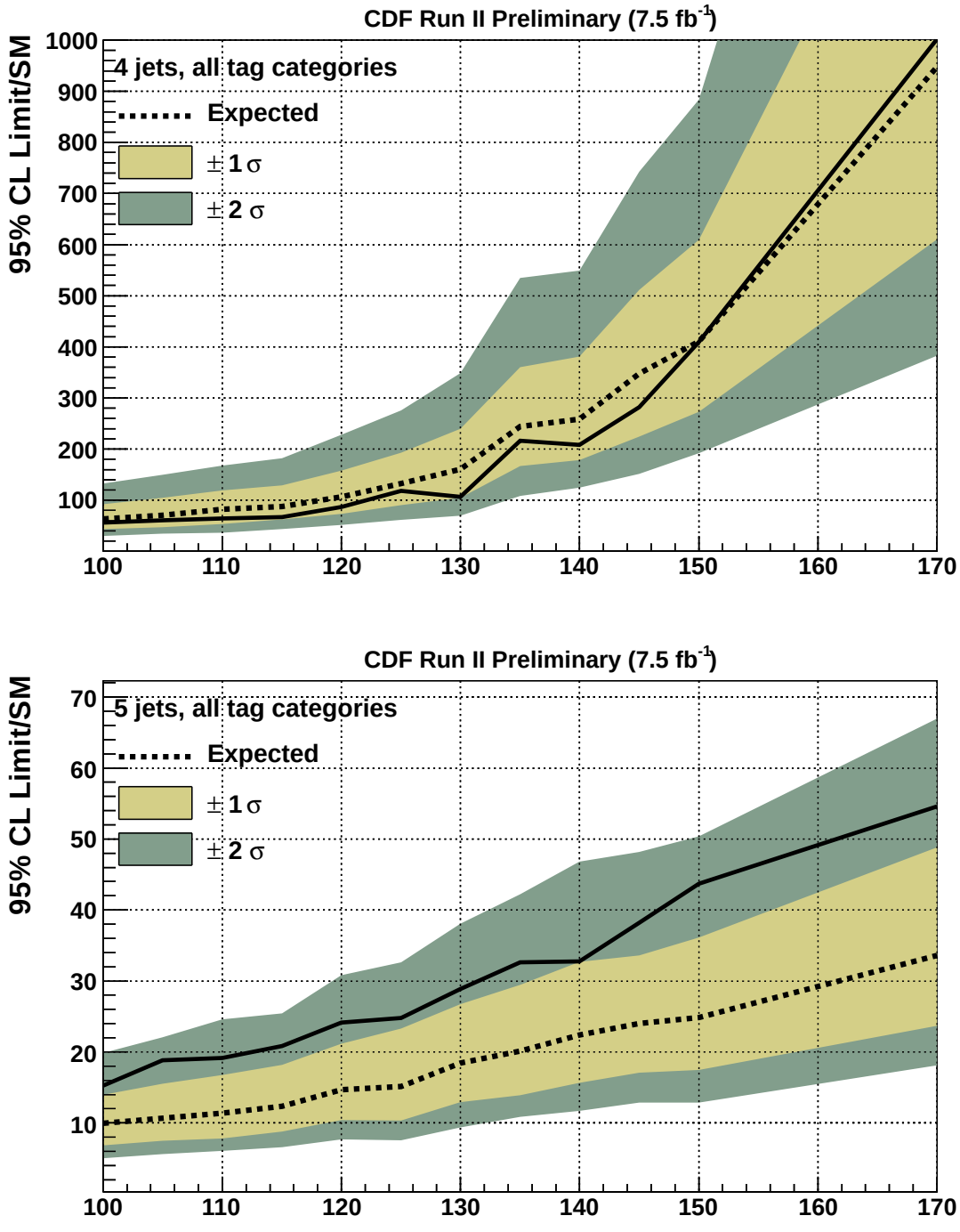

Figure 7: Expected and observed limits for the 4 jet bin alone and for the ≥ 5 jet bin alone.

References

[1] The CDF Collaboration, Search for the Standard Model Higgs Boson Production in Association with a W Boson using 5.7/fb. CDF Public Note 10239.

[2] The CDF Collaboration, Measurement of the $t\bar{t}$ Production Cross Section in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV using Lepton + Jets Events with Secondary Vertex b-tagging. CDF Public Note 7138.

[3] Enrique Palencia, Measurement of the $t\bar{t}$ Production Cross Section in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV Using Lepton+Jets Events in the CDF Detector at Fermilab. Ph.D. Thesis, CDF Public Note 8772.

[4] M.L. Mangano, M. Moretti, F. Piccinini, R. Pittau, A.D. Polosa, ALPGEN, a generator for hard multiparton processes in hadronic collisions. `arXiv:hep-ph/0206293`.

[5] Torbjorn Sjostrand, Stephen Mrenna, Peter Skands, PYTHIA 6.4 Physics and Manual,
`arXiv:hep-ph/0603175`.

[6] J. Alwall et.al., MadGraph/MadEvent v4: The New Web Generation, JHEP 0709 (2007).

[7] Thomas Junk, Confidence Level Computation for Combining Searches with Small Statistics. `arXiv:hep-ex/9902006`.

[8] D. Acosta et al., Nucl. Instrum. Methods, A494, 57 (2002).